INTERNSHIP REPORT

Developing a baseline for the bovine embryo prediction RAMP data challenge

22 JUNE 2022

JACOBO RUIZ OCAMPO

SUPERVISED BY:

FRANÇOIS CAUD

I. Introduction

My internship took place in the MIND team (Previously known as PARIETAL) at INRIA's Saclay centre. I worked mainly with François Caud (DATAIA) and to some extent with Alexandre Gramfort (MIND).
The MIND team field of study is models and inference for neuroimaging. François Caud's work with DATAIA focuses mainly in developing RAMP, a data challenge platform.

RAMP (https://ramp.studio) allows for the creation of data challenges which are competitions in which participants must work with a given dataset and come up with the best pipeline (data preprocessing and machine learning model) to solve a specific problem related to the dataset. Multiple platforms allow for the creation of such competitions however what makes RAMP unique is the fact that the participants submit their code instead of the predictions and its collaborative nature: after the competitive phase ends, participants are encouraged to share their solutions with each other in order to come up with an even better solution to the problem.

The main objective of the platform is to allow scientist that are working on a problem that could potentially be solved with machine learning, to rewrite their problem as a data challenge so that participants find the best solution to the problem while learning about machine learning.

II. Scientific problem, goals challenges and results

My work during my internship can be separated in two stages. The first stage lasted 2 weeks, I worked with the RAMP environment and acquired knowledge necessary for the second stage of my internship. During the second stage of the internship I worked on the development of a data challenge.

Phase 1: weeks 1-2

As mentioned, during the first stage I got familiar with RAMP studio and used most of my time reading documentation and learning new concepts necessary to tackle the data challenge in the second stage.
I started by reading the documentation available in RAMP Studio and studying some of the existing challenges available in the website by pulling the notebooks from their GitHub.
By the end of the first week I was able to install the RAMP server in my computer, managed to launch the frontend (the website) in local and deploy multiple challenges in the local server. The latter implies that through my local website I was able to participate in a local challenge: I could submit my solution and a score would be assigned to my solution by the server displaying it in a leaderboard.
In order to achieve this my main challenge was to understand and learn how to use conda virtual environments and manage conflicting dependencies.

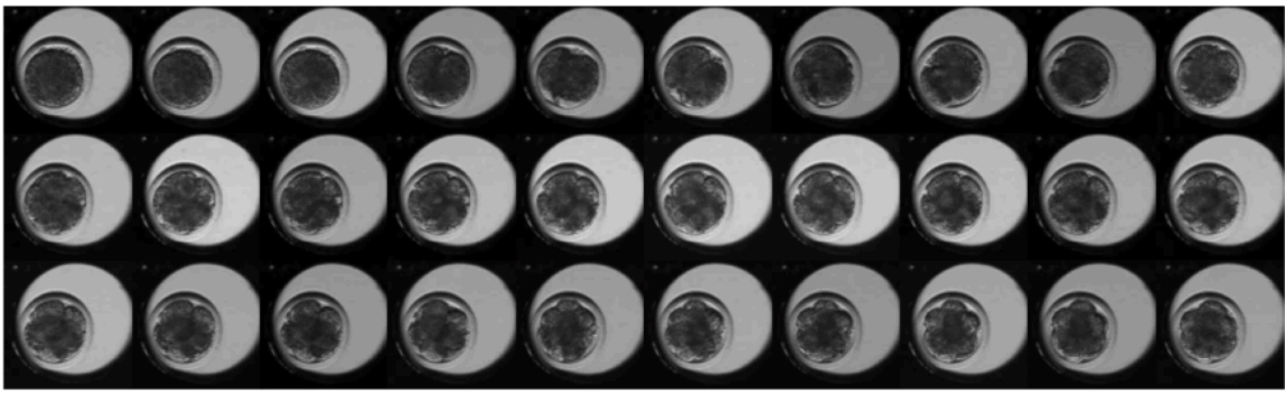
During the second week I got introduced to multiple data challenges and as all of them used neural networks in their models to solve problems I had to learn from scratch how they worked. I read and learnt from multiple sources but the main reliable source that I studied to get a deep understanding of the subject was a YouTube video series by the channel Machine Learnia that made me theoretically and practically understand a neural network by coding it in python and understanding the mathematics behind the forward and backward propagation mechanisms that are central to neural networks.

As well as learning the theory behind neural networks I had to learn how to use Keras, a python library built on top of Tensorflow which is the main library that I used to easily use neural networks in Python.

Phase 2: weeks 3-7

1. Defining the problem

During this phase I worked in the data challenge “bovine embryo survival prediction”.
It is important to understand the challenge and the dataset in order to understand my work. The dataset for this challenge consisted of 277 videos of bovine embryos (177 observations in the training set and 100 observations in the test set). Each video had 300 frames and each frame had a dimension of 250x250 pixels.



Ground Truth= A

Figure 1: 30 frames of a video of the dataset

Each video represents the evolution of an embryo between the 1st and 4th day after fertilisation, each frame is a snapshot of the embryo taken every fifteen minutes and the time at which the snapshot was taken is available to the participants. The dataset was provided by the research institute INRAE.

The problem to solve was a classification problem, the challenge is to classify between 8 possible classes denoted from A to H, A being alive and H being dead. No further information was available at this time about the differences between classes.

In addition to this, the challenge expects the participants to make 11 different predictions based on 11 different timestamps:
27, 32, 37, 40, 44, 48, 53, 58, 63, 94, None.

Each prediction will only have access to videos with frames that correspond to snapshots of the embryos taken before the timestamp.

For example, the prediction for the timestamp 27 can only use the frames of the videos of the datasets that represent snapshots of the embryo taken before 27 hours after fertilisation (I usually had access to the first 20 frames for this timestamp) in order to train and test a model.

The “None” timestamp indicates that the whole 300 frames of the videos can be used.

The timestamps are interesting to study because the time when the first snapshot of the embryo was taken varies between videos.

The problem is of a real interest because if the participants are able to create a robust model that reliably predicts which embryos are going to die, the lab could just remove them and save ressources used to nurture the embryos that would otherwise be wasted.

2. Objectives of my work

My objective was to study the dataset in order to create a non existing baseline algorithm for the challenge as well as develop tools that will help future participants of this challenge to understand the problem and the dataset faster. My goal was not to build the most performing model for this challenge, but to build a baseline classifier which by definition should perform better than a dummy classifier.

3. Analysis of the problem and challenges

A dummy classifier in the context of an unbalanced dataset (where a class is overrepresented) is a deterministic model that blindly predicts that the observation presented belongs to the overrepresented class. This is only true because a random dummy classifier would in theory have a probability to correctly classify an observation of one in eight classes (1/8 = 0.125) while in the case of the test set a deterministic dummy classifier that only predicts the overrepresented class would have an accuracy of 0.19.
My goal was therefore to train a model with the train set and test it with the test set while having a better average performance than the dummy classifier.

The first challenge was that to solve the problem at hand I had to be able to use a machine learning model that took videos as input. With the help of my internship supervisor, François Caud, I figured that the best way to begin was to simplify the challenge: I was going to create a simpler dataset where each observation was the last frame of the video (the 300th frame) instead of the original 300 frames. The simpler dataset was hence formed by 277 images that showed the embryo in its last snapshot. To simplify the task further I was considering at first that only one prediction was asked: a prediction with the “None” timestamp were I could use the whole 300 frames.

The performance of the different models discussed can be measured in two ways: as previously discussed with an accuracy which is just a percentage of correct classifications over all the classifications and with another metric defined by the biologists that created the data challenge called Weighted Classification Error (WCE) which is an error measured in the train, validation and test sets. As RAMP Studio uses the WCE metric my end goal would be to create a baseline algorithm with a lower error on the test set than the one of the deterministic dummy classifier (WCE= 0.76), hence a model with a lower WCE ont the test set on as many of the 11 predictions as possible.

4. Initial work produced while exploring the dataset

At the beginning of the data exploration phase I created tools that allowed me to better visualise a video without having to print the 300 images, such a tool was used to obtain the image in Figure 1. The tools that I developed are to be integrated in the Jupyter Notebook that introduces the data challenge to the participants.

5. A pipeline proposal with an image dataset using a CNN feature extractor and a Logistic Regression classifier

In this section I will discuss the challenges, proposed solutions and results produced by my work with the an intermediary model that only classifies the embryos for the “None” timestamp (all 300 frames of the video are available).

The reason behind the exploration of this first model is that intuitively one could believe that by seeing the last frame of the video one might be able to classify the videos as it shows the last known state of development of the embryo.

The model uses the access to the whole 300 frames to create a dataset where each observation corresponds to only the 300th frame of the video in the original dataset (each observation is represented by a single frame instead of 300 frames). From now on this dataset formed by the 300th frames of the videos will be referenced as the image dataset.

The task at hand is therefore to create a pipeline and a classifier that given an image dataset, it manages to outperform a dummy deterministic classifier.

I learnt that neural networks need a substantial number of observations to correctly learn the features of different classes of an image dataset, one of the most popular image recognition datasets is ImageNet, with over 14 million observations. A clear problem that arises is the fact that we only have 177 observations to train our model and 100 observations to test our model.

To solve this problem I used a technique used in another challenge by my supervisor called transfer learning. With this technique I use a model that has been trained to identify certain shapes with ImageNet and then I apply it to my dataset so it extracts the shapes (features) that it recognises.
Therefore I used one of Keras pre-trained convolutional neural network models: MobileNetV2 which was pre-trained with ImageNet as my feature extractor. The idea was to extract the shapes recognised in every image and use this array of shapes recognised (called a feature map) to train a classifier.

The pipeline is as follows:

- My training dataset is made up of 177 images, and my dataset is made up of 100 images.
- The images of both datasets are preprocessed in the exact same way: all the images are reduced to a size of 224x224 (previously 250x250) and preprocessed with the standard preprocessing necessary for transfer learning with MobileNetV2.
- The output of the model is not a prediction but a feature map for each frame, an array of 1280 filters (shapes recognised by MobileNetV2) per frame, so a 2D array of dimensions 177x1280.
- The feature maps are used to train a LogisticRegression classifier for multi-class classification.

The model (Logistic Regression) trained with the feature maps extracted with a MobileNetV2 based feature extractor had a WCE on the test set for the “None” timestamp prediction of 0.627 which is already a better WCE than the one of the deterministic dummy classifier.

6. Tackling the 11 predictions challenge

The previous model only predicts the class based on the very last available snapshot of the embryo (the 300th frame), however this is only possible if we have access to the 300 frames of the video (predicting for the “None” timestamp).

If we have access to for instance 200 frames the logic thing to do would would be to take the very last available snapshot of the embryo, that being the 200th frame in this case. Therefore, to tackle the other 10 predictions limited by their timestamps I created another 10 different models, each model defined by the following rules:

- In the train and test set used by the model, observations are represented by the very last available snapshot of the embryo (the very last snapshot taken before the timestamp). Therefore each dataset is an image dataset and not a video dataset.
- The model uses the same feature extractor and classifier as explained previously.

A reminder of the timestamps:
27, 32, 37, 40, 44, 48, 53, 58, 63, 94, None.

I trained and tested the 11 different models and plotted their respective WCE in the following figure:

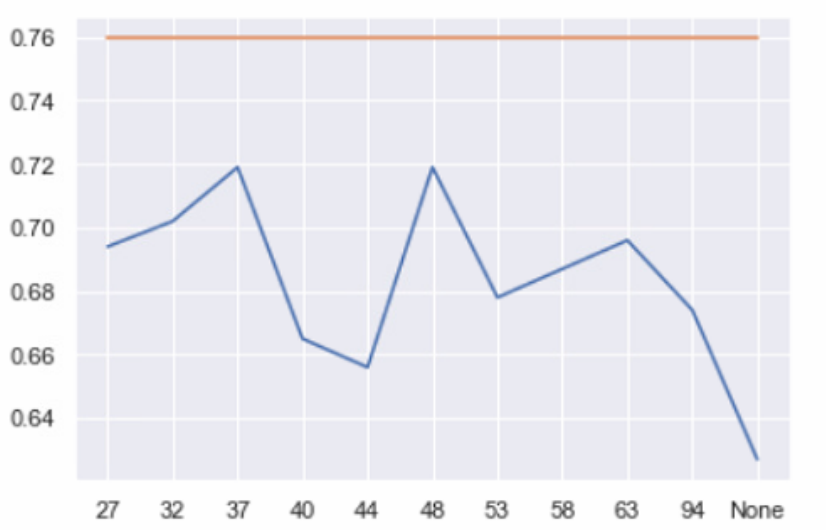


Figure 2: In ordinate the WCE, in abscissa the timestamps corresponding to the 11 models studied (48 corresponds to the model that used snapshots taken up to the 48th hour). In blue is plotted the performance of the final 11 models that represent my baseline and in orange is plotted the performance of the deterministic dummy classifier.

In the figure 2 we can see an overall trend of a lesser WCE as we take images closer to the end of the video. It is interesting to see that every single model performs better than the deterministic dummy classifier (WCE= 0.76) and the best performing model has a WCE of 0.627 (the one for the None timestamp).

7. Further work

My internship was quite limited by time and I would have liked to study the following even further:

- I learned that data augmentation techniques like rotating the images helped boost performance of the models and is a path towards a better baseline pipeline that I would have liked to immediately explore.
- I researched about new models that could use a 3 dimensional feature map using the temporal dimension of the sequence of frames to include this critical trait and then feeding this feature map to a RNN or a Transformer. This seems the natural way to proceed in terms of model exploration as my solution ignores the fact that the data set is made of videos.
- I reduced the number of classes to 2 by grouping classes A,B,C,D together and E,F,G,H together to explore the results of a binary classification. A further discussion with biologists helped enlightening the mysterious definitions that separate the 8 different classes. I synthesised the discussion in an annex.

Some of this techniques helped improving my existing baseline but some of them, specially the 3D feature maps where a bit too technical for my current skills as they required me to use advanced neural network architectures.

Remaining work: weeks 8-9

During my remaining time my objective is to write this baseline that I have created in the form of a python script that can be submitted to RAMP and then it can be used by the rest of the development team to create a better baseline for the official data challenge.
When the script will be uploaded the score obtained by the baseline will be visible in the leaderboard of the challenge in the Ramp Studio website:

(https://ramp.studio/events/bovine_embryo_survival_prediction_open_2022/leaderboard).

III. Feedback

This year's course in machine learning was specially useful as it taught me the bases of how a machine learning pipeline works and the theoretical bases of the unit at the heart of neural networks: the perceptron (a single neuron). Last year's course in introduction to machine learning was also very helpful as it introduced me to the world of data challenges.

I specially enjoyed being given such a challenging task to accomplish without having much knowledge in the area, I enjoyed being forced to rapidly learn and adapt myself to new things (mainly talking about the Keras python library).

I would have liked to have more time to learn and explore the world of the video classifiers which are more complicated but required a lot of time to understand and learn how to use. I am however looking forward to learning more about Neural Networks next year in the Artificial Intelligence Master.

IV. Acknowledgements

I specially thank Sarah Cohen-Boulakia, Sylvain Conchon, François Caud and Alexandre Gramfort for helping me make this internship possible.

IV. Annex: Conversation with the biologists that created the challenge

My tutor and I met with the biologists and other researchers in charge of creating the challenge in order to further understand and seek clarification on how the 8 classes are defined.

The meeting took place by the end of my internship and I wasn't able to use the knowledge gained due to the time constraint but here I present my notes of the meeting regarding what was discussed about the differences between the different classes.

A: embryo evolves constantly as time passes on time, at some point the cells divide in a an a-synchronic manner (meaning 1 cell is visible, then two, then three).

B: class hardly discernable from A except for the fact that the division is synchrone (meaning one cell is visible, then two then four).

C: starts as a normal division but then "loses rhythm" and has a difficulty to divide (longer time taken to divide).

D: Begins very badly but ends up living (abnormal division, looks like it is dying).

E: starts as embryos of class C but ends up badly (dies).

F: starts as embryo of class B but dies after the third cycle of cellular division.

G: Embryos that stay alive but growth stagnates (same number of cells for a prolonged period: stops dividing, even if cells move around)

H: at some point it explodes and dies.

Based on this it is obvious that the model proposed during my internship would have quite a difficult time differentiating between class E and F for instance as both of this classes need a model that studies both the beginning and the end of the video.