Project 1: Learn parasitology and diagnose with Deep Learning

Objectives:

- Train explainable Deep learning models in the same manners a human domain expert teaches his students.
- Make models more robust, generic, and explainable.
- Train models that are able to learn with few-shot learning, on which we can apply transfer learning from other domains/datasets
- Publish a Parasitology dataset to help future research such as faster diagnosing of infectious diseases (e.g., Malaria transmitting parasites when expert advice is not present).
- Produce deep learning model outcomes that provide explanations alike to the domain expert, and thus, that is right for the right reason[1], and not learning data bias.

Tasks:

- Collect a small dataset for parasite aspect identification and classification according to biology taxonomies
    - Collect data (using perhaps a web scraper such as Python Beautiful Soup from https://www.cdc.gov/ (investigate best way on their site) from a given 5 test classes. Start collecting attributes, and then building the annotation tool will help expert annotate with bounding boxes if needed, at a second stage. Example for Malaria images.
        - Idea to explore: Python: Create Your First Web Scraper with ScrapingBee API and Python(cloud-based scraper can also be applied to other languages).
    - [Future work, next year, when data is available] Collect data from University of Granada Parasitology course and perform dataset augmentation
- Use existing tool to set up a dataset annotation tool to label the images
- Perform multi-level annotation with expert indications, material, and taxonomies
- Format and publish the dataset
- **ATTRIBUTES LIST:** Building a multiview DNN model to classify, with a single model, each image according to the biological taxonomy. At least the following **attributes** for a systematic classification/taxonomy will be annotated:
    - *Phylum*
    - *Class*
    - *Specie*
    - *Form*
    - *Sample* (blood, fecal, food)
    - Others: host (environment, human, animal), sex, and other discriminative features for consistent outputs.

[1] Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations Andrew Slavin Ross, Michael C. Hughes, Finale Doshi-Velez

- Building a generative model for data augmentation
    - Evaluation: Can we condition on the different aspects and taxonomic properties of an individual that still produce coherent and non contradicting samples and predictions? I.e. could a single model as the 12 in 1 perform domain-expert like classifications according to different taxonomies?[2]
- Augmenting the original baseline model for classification with scrabble drawings/pattern geometric simplifications in order to find if feeding these in the training or in the inference time, they help the model identifying the discriminating characteristic for which a data point can be properly classified/generated instead of misclassified/generalizing to other condition. E.g. using an embedding representation for sketches from [6].
- Models evaluation and scientific publication report

IMAGE ANNOTATION (I.e. labelling) TOOLS:

- **For image annotation you may be aware of via from vgg (Julien uses): https://www.robots.ox.ac.uk/~vgg/software/via/via_demo.html**
- https://www.sicara.ai/blog/2019-09-01-top-five-open-source-annotation-tools-computer-vision
- https://towardsdatascience.com/annotate-your-image-using-online-annotation-tool-52d0a742daff
- https://hackernoon.com/the-best-image-annotation-platforms-for-computer-vision-an-honest-review-of-each-dac7f565fea

REFERENCES

[1] https://github.com/cs-chan/ArtGAN
[2] Explaining digital humanities by aligning images and textual descriptions
M Cornia, M Stefanini, L Baraldi, M Corsini, R Cucchiara Pattern Recognition Letters 129, 166-172
[3] CausalNex is a Python library that uses Bayesian Networks to combine machine learning and domain expertise for causal reasoning
[4] Accessible cultural heritage through Explainable Artificial Intelligence. N Díaz-Rodríguez & G Pisoni. 2019. https://www.dropbox.com/s/uielmu43xdilidx/UMAP_PATCH_2020.pdf?dl=0
[5] Human and Automatic Detection of Generated Text https://arxiv.org/abs/1911.00650
[6] A Neural Representation of Sketch Drawings https://arxiv.org/abs/1704.03477
[7] Parasitology potential imaging software to test on our algorithms/annotation tool:

[2] Classifying objects by their discriminative features, as experts do, not by learning bias in the data.

DeepImageJ: A user-friendly plugin to run deep learning models in Image. JE Gómez-de-Mariscal, C García-López-de-Haro, L Donati, M Unser, ...bioRxiv, 799270
[8] Bayesian neural networks that learn correlation provide more robust uncertainties than a single model via an ensemble of bayesian NNs and shows that explicitly incorporating domain-specific knowledge both improves performance and provides additional insight by inferring the covariance of the retrieved atmospheric parameters. besides better accuracy, the Bayesian technique offered something equally as critical: it could tell the scientists how certain it was about its prediction. “In places where the data weren’t good enough to give a really accurate result, this model was better at knowing that it wasn’t sure of the answer, which is really important if we are to trust these predictions,”
[9] "Estudio morfológico y biológico de giardia SPP: aspectos epidemiológicos de la giardiasis humana y animal en la provincia de Granada" VD Sáez -
[10] P Gijón Robles 2013 DIAGNÓSTICO DE PARÁSITOS EN HECES: COMPARACIÓN DE DOS TÉCNICAS DE CONCENTRACIÓN.

Application inspired references: E.g. XAI for COVID:

- DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images. Karim et al.: https://lnkd.in/eNWQw7P
- SurvLIME-Inf: A simplified modification of SurvLIME for explanation of machine learning survival models LV Utkin, MS Kovalev, EM Kasimov - arXiv preprint arXiv:2005.02387, 2020
- Grand challenge on Pathology Visual Question Answering and invite participations. The task is to develop models to answer questions about the contents of pathology images. There are 32799 questions from 4998 pathology images. Is it possible to develop an "AI Pathologist" to pass the board-certified examination of the American Board of Pathology (https://www.abpath.org/index.php/taking-an-examination/sample-examination-questions <https://www.abpath.org/index.php/taking-an-examination/sample-examination-questions>)? To achieve this goal, we launch this medical visual question answering (VQA) challenge where deep learning models are to be developed for interpreting pathology images and answering questions about the image contents. The task is VQA on pathology images: given a pathology image and a question, the model needs to give the correct answer. We have collected a dataset containing 4,998 images and 32,799 question-answer pairs. Half of these questions are open-ended (why, what, how, where, etc.) and the other half are “yes/no” questions.
  The data is available at https://github.com/UCSD-AI4H/PathVQA It contains 4,998 pathology images and 32,799 question-answer pairs. Half of these questions are open-ended (why, what, how, where, etc.) and the other half are “yes/no” questions. We

provide an official split of training, validation, and test. The details of this dataset are described in this preprint. https://arxiv.org/abs/2003.10286 For evaluation and challenge rules, please refer to https://pathvqachallenge.grand-challenge.org/

Protocol to gather and annotate the data:

1. Search the species in The CDC DPDx site using the A-Z index.
2. Annotate the attributes in the attributes list and save both image filename and attributes in a csv text file.
    a. When unsure of the forms of the parasite, google the species, e.g.: *myxobolus cerebralis* and click on Images.
    b. Then see the forms of the parasite, some show cysts, eggs, worms or other scientific forms. This will help you give an idea of the different parts in which the parasite life cycle should be important to identify by our system, and therefore, important phases to collect data on. Try to begin with as many images as possible of each form (10 if you can, or the max available (from CDC only)). The form of the parasite cycle that is most important is the form in which the parasite passes through the human so we can do diagnosis on it (normally parasites are living after ingestion of food (meat or fish[3]) in the liver, and it can be screened with imaging. However they can also be diagnosed in fecal samples after infection, also with microscopy or regular images.
        i. Therefore, try to prioritize the collection of images from the parasite phase that is the one living in the human being if you are unsure. In order to find the phases, once you find the parasite through clicking on the A-Z index of the CDC site[4], e.g for Trichinellosis, you will end up here: https://www.cdc.gov/dpdx/trichinellosis/index.html
3. Then click on to see the images and try get from different cycles. You normally see the life cycle in the first tab (Parasite biology) in a drawing and can see the name of the cycle inside the human, and other cycles outside. It is also important to collect data from those parasites in the original source so we could later develop an app that tcould screen food for poisoning in e.g. meat or fish. They for instance cause a whirling disease in the fish that eats them (see final figure and other phases in the life cycle here[5]).
4. Once you found the phases or forms of the parasite, e.g.: Larva, encysted larva, etc. These will be features to annotate in your images. Try to annotate everything you can that is available in CDC from the list in the first page.

---

[3] https://www.google.com/search?q=myxobolus+cerebralis&safe=active&rlz=1C5CHFA_enUS847US848&sxsrf=ALeKk03miRph9jGvyT3dIdtVS6UuAw2TkA:1590334317952&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjdweGA6czpAhXzAWMBHa_PBO8Q_AUoAXoECBcQAw&biw=1933&bih=1245

[4] CDC DPDx site https://www.cdc.gov/dpdx/az.html

[5] https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/animal-plant/aquatic/field-guide/4th-edition/finfish/whirling-disease.pdf

Recap of families to collect data in all its life cycles:

- ARTRÓPODOS (like bedbugs, ticks)
- TREMATODOS (look like a leaf)
- CESTODOS (look like a lace, a flat strip with segments on it)
- NEMATODOS (look is cylindric, a pencil like shape but microscopic)
- PROTOZOOS (unicellular)

The taxonomy of parasites to find images and attribute label data are the following (note some notation is in spanish, please translate (Quistes = cyst, etc):

CTA (Food science and technology) Subject:
Entamoeba histolytica
o Trofozoítos
o Quistes (cyst)
Giardia lamblia
o Trofozoítos
o Quistes
Iodamoeba bütschlii
o Quistes
Balantidium coli
o Trofozoítos
o Quistes
Entamoeba coli
o Quistes

Trichinella spiralis (adulto)
Trichinella spiralis (larvas en musculo estriado)
Echinoccocus granulosus (adulto)
Echinoccocus granulosus (quiste hidatídico)
Taenia spp. (cisticerco)
Sarcocystis spp (tejido muscular)
Toxoplasma gondii (quiste)
Larvas de mosca (Sarcophaga y Calliphora, larvas II y III)

o Fasciola hepatica:

Adulto
Huevos
Miracidio
Redia
Cercaria
Metacercaria
o Dicrocoelum dendriticum
adulto
o Clonorchis sinensis
Adulto
Huevos
o Diphyllobothrium latum
Huevos
o Hymenolepis nana
Huevos
o Hymenolepis diminuta
Huevos
o Taenia spp.
Huevos
o Toxocara canis
Huevos
o Ascaris lumbricoides
Huevos

NHD (Human Nutrition and Diet) subject

Trematodos
- Fasciola hepatica: Adulto
Huevo
Metacercaria
- Clonorchis sinensis: Adulto
- Dicrocoelium dendriticum: Adulto
Cestodos
- Taenia sp.: Escólex
Anillo inmaduro
Anillo maduro (sexuado)
Anillo grávido
Huevo
Cisticerco
Estrobilicerco
- Hymenolepis nana: Escólex y anillos

- Echinococcus granulosus: Quiste hidatídico (sección)
- Diphyllobothrium latum: Anillos

Fasciola hepática (adulto)
F. hepatica (huevo)
F. hepatica (metacercaria)
Clonorchis sinensis (adulto)
Dicrocoelium dendriticum
Taenia sp. (escólex y anillos)
Taenia sp. (huevos)
Taenia sp. (cisticerco)
Taenia sp. (estrobilicerco)
Hymenolepis nana (escólex y anillos)
E. granulosus (quiste hidatídico)
Diphyllobothrium latum (anillos)

- Trichuris trichiura: Adultos
Huevos
- Trichinella spiralis: Larvas en músculo estriado
- Ascaris lumbricoides: Huevos
- Enterobius vermicularis: Hembra
- Ancylostoma sp: Adultos
- Strongyloides sp: Larvas rabditoides

Trichuris trichiura (adultos) 4X 4-10X
T.trichiura (huevos) 10X 40X
Trichinella spiralis (larvas) 4X 10-40X
Ascaris lumbricoides (huevos) 10X 40X
Enterobius vermicularis (hembra) 4X 10-40X
Ancylostoma sp.(adultos) 4X 10-40X
Strongyloides sp. (larvas rabditoides)
Larvas de Anisakis

- Entamoeba histolytica: Trofozoítos
Quistes
- Giardia lamblia: Trofozoítos
Quistes
- Toxoplasma gondii: Quistes tisulares
- Plasmodium falciparum: Formas sanguíneas (en anillo, esquizontes y gametocitos)
- Phylum Apicomplexa: Ooquistes

E. histolytica (trofozoítos, quistes) 40X 100X
Giardia lamblia (trofozoítos teñidos) 40X 100X
G. lamblia (quistes en fresco) 40X 40X
Toxoplasma gondii (quistes tisulares) 10X 40X
Phylum Apicomplexa (ooquistes de
coccidios en fresco)
Plasmodium falciparum (forma en
anillo, esquizonte, gametocito)

- Tribolium confusum: Adulto
Larva
- Oryzaephilus sp: Adulto
- Calliphoridae: Larvas
- Sarcophagidae: Larvas
- Tyrophagus putrescentiae: Adultos

Tribolium confusum 4X 4-10X
Oryzaephilus sp 4X 4-10X
Larvas de mosca 4X 4-10X
Tyrophagus putrescentiae

Protocol of food parasitology (CTA):

Huevos de Ascaris lumbricoides, Toxocara canis, Taenia spp, quistes de Giardia lamblia y Entamoeba coli.

Quistes de Giardia lamblia.

Quistes de coccidios.

Huevos de Taenia sp. y Toxocara canis.

Observacion de caja de preparaciones con:
Entamoeba histolytica
o Trofozoítos
o Quistes

Giardia lamblia
o Trofozoítos
o Quistes
Iodamoeba bütschlii
o Quistes
Balantidium coli
o Trofozoítos
o Quistes
Entamoeba coli
o Quistes

Trichinella spiralis (adulto)
Trichinella spiralis (larvas en musculo estriado)
Echinoccocus granulosus (adulto)
Echinoccocus granulosus (quiste hidatídico)
Taenia spp. (cisticerco)
Sarcocystis spp (tejido muscular)
Toxoplasma gondii (quiste)
Larvas de mosca (Sarcophaga y Calliphora, larvas II y III)

o Fasciola hepatica:
Adulto
Huevos
Miracidio
Redia
Cercaria
Metacercaria
o Dicrocoelum dendriticum
adulto
o Clonorchis sinensis
Adulto
Huevos
o Diphyllobothrium latum
Huevos
o Hymenolepis nana
Huevos
o Hymenolepis diminuta
Huevos
o Taenia spp.
Huevos

o Toxocara canis
Huevos
o Ascaris lumbricoides
Huevos

Legend notes (these parasites are based in the Food parasitology course from University of Granada):

- sp.: species; spp.: latin nomenclature for species (plural).
- Discard the observation details (e.g. numbers such as 40X, etc, they are for real lab).
- Also discard some repeated parasite names, but account for all the forms in which they can appear in a python dictionary that you will need to produce a taxonomy for later in the dataset (and overleaf report).
- Example of fish to observe some anisakidos: Bacaladilla (Micromesistus poutassou). Collecting images of parasites in foods would be very valuable for diagnosis of food quality. Assessing if DNNs can diagnose parasites in water, fruit and vegetables is one of the final objectives of teaching parasitology.
- Helmintos: transform the head of the animal (fish?) that should be flat into one that looks like a dog due to the deformations that the parasite produces inside.

Future work: I will add more from slides, but these are all a good set to begin with collecting images and attribute (tabular) based annotations.

The objective is making the DL model being able to explain the output of their classification and also align their prediction with an explanation that is based on being able to pinpoint the feature leading to such classification in the same manner as parasitologist do, i.e., signaling the feature that is pathognomonic, i.e. the characteristic in the image that is unique to that specie only. Our task is to demonstrate whether a DNN is able to do that.

Add Team weekly updates below:

22-05-2020

Here is the summary for this week.

- We searched for malarial parasite data and found https://www.cdc.gov/dpdx/malaria/index.html?fbclid=IwAR16g4y0BbehDX3zMiWFgUwYCk0UE5F43Hp1nVxT4z5jlS4hPo11ycFQRTg (in the image gallery section).
- We scraped the webpage and downloaded the microscopy images and the corresponding description given with each image.

This is our plan for next week.

- Annotate the data
- Build a baseline model for classification of parasites into their correct species.

Ideas that we need feedback on. We request you to provide your valuable feedback.

- As only a few cells are infected in a given image, before annotating, we are thinking of segmenting the particular cells with parasitic infection and then upscaling the resulting images. By doing this we think it will be easier for the deep learning model to identify the parasitic infection pattern. After that, we will annotate the image.

29-05-2020

Here is the summary for this week:

- Downloaded images for most of the species mentioned in the doc.
- Started annotating the data.

Here is the plan for the next week:

- Finish annotating the data.
- Build a basic model for classification.

Points that we need feedback on. We request your valuable feedback:

- While scraping the website, along with the images, we downloaded the metadata and saved it in the form of a csv file. The saved file has columns such as Image ID, phylum, class, genus, species. An image annotator-if we use it-will label the images and save the labels in the form of a csv file. Since we already have a csv file containing data such as Image ID, class, genus, species, and since we just have to label the images and not draw bounding boxes, we are thinking that instead of using an image annotator and again annotating the fields which we already have (phylum ,class, genus,species), we can add columns such for other attributes in the csv file and label in the file itself. This way we can save some time.

06-06-2020

Here is our summary for this week:

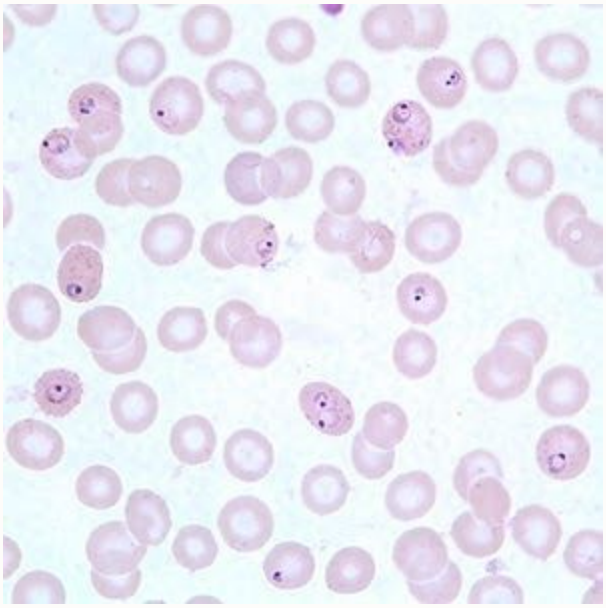
- Finished with csv files containing the fields for annotation.
- The fields for annotation currently are for Phylum, Class, Genus, Species, Form and Sample.

Plan for next week:

- Annotate the fields, encircling the parasites (pointing at the features that make them distinct) with the help of VGG Image Annotator.
- If finished with annotating, try the baseline model.

Points that we need feedback on. We request your valuable feedback:

- In the image that I have attached with this below, there are many cells that are infected with the parasite. Should we encircle and annotate all the infected cells (by drawing box around them) or encircling just one infected cell is sufficient.



12-06-2020

Here is our summary for this week:

- Developed a baseline model for classification.
- This model consists of a pre-trained CNN (ResNet-101) base with a fully connected layer having parallel outputs, where each output is for a particular feature (phylum, class, species, sample or form).

Plan for next week:

- Train the model.
- Try different pre-trained models.

Point that we need feedback on. We request your valuable feedback:

- In addition to the baseline model, we are thinking of developing another model that uses CNN to create feature embedding of the image. Then we will treat this feature embedding similar to a word embedding and use a Skip-Gram model to predict the context words (context words in this case are phylum, class, species etc.) given the feature embedding.

19-06-2020

Here is our summary for this week:

1. Trained on baseline model. The model uses pretrained cnn that extracts the feature embeddings, has a trainable layer and after that splits into five outputs with softmax activation for five attributes.
2. Tried another model that has a transformer block before the output layers.
3. The model in 2 performed better than in 1. But it still needs improvement.
4. Experimented with architectures such as lenet, vgg16, resnet101,resnet50 as pretrained feature extractors.

Plan for next week:

Try more cnn architectures and try to improve the performance of the model.

26-06-2020

Here is our summary for this week:

- Experimented with different architectures.
- Further processed the images to remove unwanted features in the image (like some images had arrows and text, so we removed them).

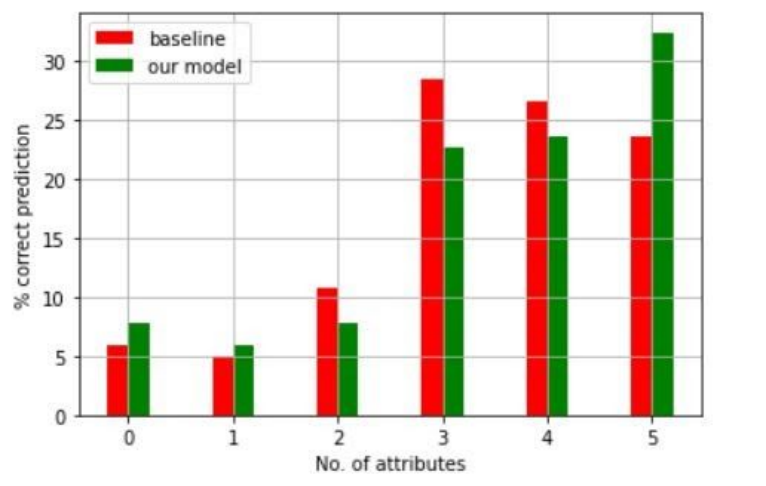
Plan for next week:

- Try to improve model performance.

Point that we need feedback on. We request your valuable feedback:

We have a limited amount of data. For many organisms, we have diverse images with many having just one or two images of a particular type. It is very likely that during evaluation (if test data is selected randomly), some types of images may appear that have not been trained in the network. Therefore, they are likely to be predicted incorrectly even if we augment training data (by translating, rotating etc.) because these images will still not appear in training data. The training data is also limited. Since the number of labels is large, this problem is significant. We request you to suggest a way around to this.

Attached below is the plot for model performance. The plot indicates how many times n number of attributes are predicted correctly in test data.



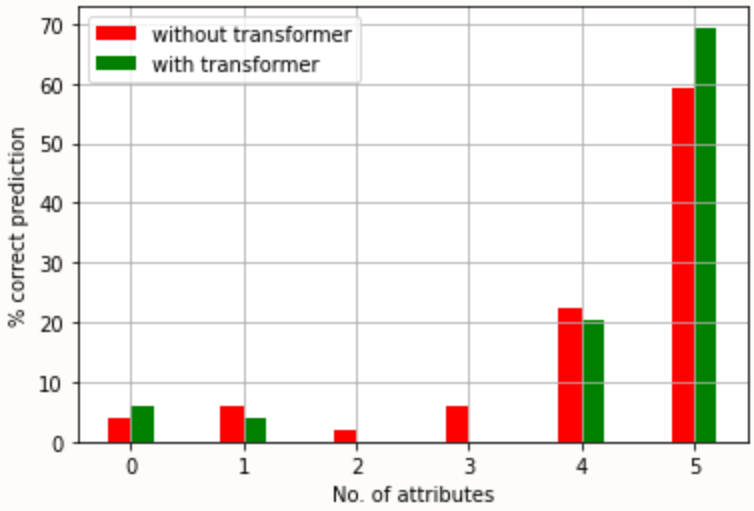
06-07-2020

Summary for the last week:

- Filtered the data such that the labels with less than 10 samples were removed. The model building and prediction was done only with this data.
- Tried image augmentation using smote, as well as by doing random changes like zooming, rotation, changing brightness but none of that seemed to work.
- Also tried to impose loss weights, class weights to the model but it did not work as well.
- Tried different number of the layers and different positions of the transformer block.

Plan for this week:

- See if the model can be made better.
- Work on the annotation tool.



11-07-2020

Summary for this week:

- Worked on the model. Tried changing making changes within the transformer block but it did not work.
- Progressed with gradcam.
- Started working on the annotation tool.

Plan for next week

- Finish the gradcam part
- Work on the annotation tool

19-07-2020

Summary for this week:

- Implemented gradcam.
- Worked on annotation tool. It is almost complete

Plan for next week:

- Finish the annotation tool