Project Ideas

My idea for this project is divided into 2 segments.

SEGMENT 1

The first segment is about **effective model pre-training strategies** for tasks related to drug discovery. Broadly speaking, I have identified three different kinds of pretraining methods. First, is a weakly-supervised training procedure (as proposed in the ChemNet paper) where we train our model on a Larger, similar-domain dataset (such as chEMBL) and then fine-tune on a smaller dataset (such as Tox-21). Second, we have the BiG Transfer pretraining by Google AI (https://ai.googleblog.com/2020/05/open-sourcing-bit-exploring-large-scale.html) where they claim to have created a general-purpose pretraining for all tasks related to computer vision (even including specialist tasks like Medical Imaging). Third, we test the vanilla imagenet-pretraining which is widely adopted to all vision-tasks but has been highly debated as well. ResNet50 architecture can be used a a common architecture for all the experiments.

In this study, we can also answer questions like ***"do different transfer learning strategies in fact result in any representational differences?***

this can be done by quantitatively studying hidden representations like (SV)CCA [http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-understanding-and-improvement]

When using BiT, try experimenting with both 1-shot (1 sample per class) and 5-shot (5 samples per class) methods and with wider-variants of ResNet.

- [ ] chEMBL pretraining
- [ ] BiG Transfer
- [ ] ImageNet Pretraining
- [ ] Noisy Student Pretraining (Self-supervised learning)

SEGMENT 2

▼ Segment 2 is about evaluating different model architectures for drug-discovery related tasks. The paper ***"Transfusion: Transfer Learning for Medical Imaging"*** talks about how the imagenet related models (which are also de-facto for all vision related tasks) might be over-parameterized for tasks involving smaller datasets (which is common in Cheminformatics). In this segment, we evaluate three different types of architectures. Firstly, basic image-classification models which follow the traditional C-B-R progression. These models can be further divided into categories "Tiny", "Small" and "Large" based on their size. The second category of model architectures is those especially designed for ImageNet classification (ResNet, DensNet and EfficientNet). The final category of models is those which are neither traditional nor specialized for ImageNet such as the Capsule Networks (which take hierarchical relationships into account which is important in chemical compounds). If there is a huge gap in the model performances, we can look into interpretation methods like GradCAM, Integrated Gradients (see tf-explain library for a simple implementation of these methods) to look further into the model decision-making process.

The evaluating criteria for this study can be basic classification metrics like accuracy, f1-score, AUC-ROC.

- [ ] Evaluate different architectures for Drug-discovery related tasks
- [ ] Basic Models with CBR progression (tiny, mid, large)
- [ ] Models for Imagenet (ResNet, DenseNet, EfficientNet)
- [ ] (Kinda) specialized models like CapsuleNet
- [ ] Use interpretation methods to see what the models are looking while predicting. (tf-explain)
- [ ] While doing data-augmentation, try standard methods and MixUp as a comparative study

Important Resources for this work

Papers

- [ ] Big Transfer
- [ ] Capsule Networks for Protein Structure Classification and
- [ ] Domain Adaptive Transfer Learning with Specialist Models
- [ ] Inductive transfer learning for molecular
- [ ] Transfusion - Understanding Transfer Learning for Medical Imaging

Tools

- [ ] (SV)CCA - [https://github.com/google/svcca, https://ai.googleblog.com/2017/11/interpreting-deep-neural-networks-with.html]
- [ ] tf-explain - [https://tf-explain.readthedocs.io/en/latest/]
- [ ] BiG Transfer - [https://github.com/google-research/big_transfer, https://ai.googleblog.com/2020/05/open-sourcing-bit-exploring-large-scale.html, https://blog.tensorflow.org/2020/05/bigtransfer-bit-state-of-art-transfer-learning-computer-vision.html]