# Multimodal representation learning of brain pathology

**Abstract**

We propose a method for learning features of pairs of magnetic resonance images of human brains and their respective medical report. This representation can be used to train a classifier, or to match other samples with similar features by computing a similarity score. We train a deep learning model with a contrastive loss, which learns to match image-text pairs. We depend on well-established state-of-the-art methods to construct our architecture. Finally, we examine the obtained results, demonstrating that although the suggested method is not yet optimal, it still exhibits promising behavior that could be investigated further in future work.

**Keywords:** Deep learning, representation learning, multi-modal learning, brain pathology, magnetic resonance imaging, self-supervised

## 1 Introduction

Through automated medical imaging analysis, scientists are developing methods to ease the work of healthcare professionals in clinics or to improve the workflow of medical research by making it more efficient. Magnetic resonance imaging (MRI) protocols are used by physicians to diagnoses and research different pathologies in many organs, including the brain. Different protocols and sequences affect the type of signal that is acquired by MRI scanners, depending on what kind of information is required to look at a specific phenomena, as it reacts differently based on the scanned tissue. For example, when inadequate blood supply is flowing to a certain area, the affected cells might die with consequent tissue necrosis. The dead tissue is called infarct, which is better detected by radiologist on a MRI sequence called diffusion-weighted imaging (DWI). In particular, DWI is useful for the early identification of ischemic stroke, and to differentiate acute from chronic stroke. Early action on ischemic stroke can reduce infarcts, and the consequent brain damage or other complications. A similar discussion can be made for many pathologies and MRI sequences. In a clinical context, when such images are acquired, a radiologist would also annotate its findings in a medical report. We propose an automated machine learning method for learning a feature representation of such image-text pairs. In order to do so, we train a CLIP-like model to efficiently learn visual concepts from images and natural language semantics from

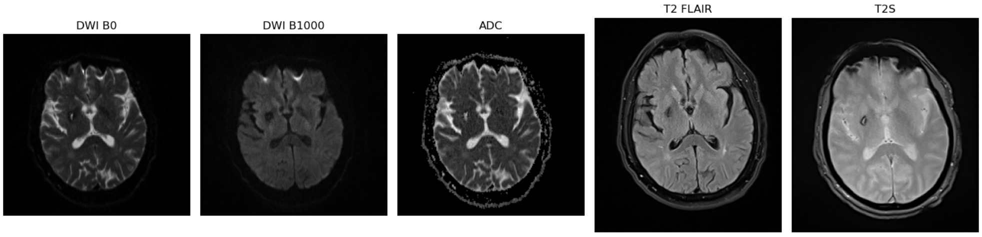

**Fig. 1** One case from the dataset and its available MRI sequences, with findings: "Chronic infarcts in the right lentiform nucleus, right corona radiata, body of right caudate nucleus posterior limb of right internal capsule. Hemosiderin deposits in the posterior limb of right internal capsule Age related cerebral atrophy with Fazekas Grade I small vessel ischemic changes. Left maxillary sinusitis."
We train only on DWI B1000, T2S and T2 FLAIR.

text. [1] Once we learn a joint representation, we can query the model to return similar images given an input sentence, or a sentence given an input image. We could use this joint representation as a feature extractor for a classification task, which aims at distinguishing normal brains from the ones affected by different pathologies.

The code of our implementation is made available through GitHub. [1]

# 2 Data

The available dataset is made by approximately 1100 patients. For each of these patients, at least three brain MRI sequences were acquired: T2-weighed (T2S), fluid-attenuated inversion recovery (FLAIR) and diffusion-weighted (DWI). All of them, or more, are required to diagnose a patient with conditions such as tumor, hemorrhage or infarct. Each image comes with a medical report written by the clinician inspecting the results of the MRI acquisition. This information is usually not present or discarded when developing machine learning methods on medical images, in both classification and segmentation tasks. The dataset owner used natural language processing to extract a diagnosis from the medical report, to classify each document as a class among tumor, hemorrhage, infarct, normal or other pathology. We use this information to sample from a balanced dataset (Table 1) the final set of images used for the training and testing of the proposed method. The selected images, after preprocessing, have 4D dimensions, the first three given by the volume of a single MRI acqusition, and the fourth is obtained by stacking the three available MRI sequences on top of the other. The three dimensions in a MRI volume are not spatially equal: the first two dimensions represent the width and height of a single image plane, while the

[1] https://github.com/aliswh/brain-CLIP

|  | total | infarct | tumor | hemorrhage | normal | others |
|---|---|---|---|---|---|---|
| training | 211 | 40 | 47 | 29 | 48 | 47 |
| validation | 24 | 5 | 6 | 5 | 3 | 5 |
| testing | 27 | 5 | 6 | 4 | 6 | 6 |

**Table 1** Distribution of samples into training splits by class.

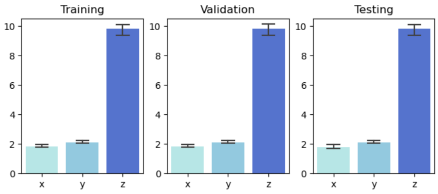

**Fig. 2** Voxel size distribution per image dimension.

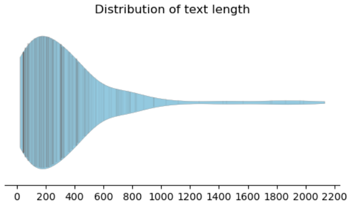

**Fig. 3** Distribution of length of medical reports in number of characters.

third provides information along the z-axis by piling consequent slices of the volume. This is due to the nature of MRI acquisition, which affects the voxel resolution of the image, that tends to be anisotropic by having less information along the third dimension. (Figure 2) The text data is extracted from the parsed impressions of each image medical report. They use medical-specific syntax and summarize the main findings of the procedure. These texts have a variable length (Figure 3) and are made by short annotations, rather than full sentences. An example of the data used can be seen in Figure 1.

## 2.1 Preprocessing and Data Augmentation

We tested the text encoder for classification on text only, and it achieved an high enough accuracy even with the raw text. Consequentely, no preprocessing was applied to the text, as we rely on the encoder for the extraction of semantically relevant features. The images were resampled and cropped to fit a size of 112 × 112 × 16 voxels. Moreover, three MRI sequences were co-registred to overlap on the axial plane, having a final image sample shape of 112 × 112 × 16 × 3. Images were augmented on-the-fly during training, with a probability $p = 0.5$ of having any of the following transformations from the TorchIO package [2]: affine transform with $p = 0.8$ (scaling, rotation and translation), Gaussian noise, flip along left-right axis and change of contrast. Parameters are available in the code repository.

# 3 Previous work

The work done for this study is based on three other well-established methods: CLIP [1], DistilBERT [3] and 3D U-Net [4]. As a model, CLIP (Contrastive Language–Image Pre-training) pre-trains an image encoder and a text encoder to predict which images were paired with which texts in a given dataset (Figure 4). This learnt behaviour is then used for zero-shot classification, which aims at predicting the classes of a new test set, even when they were not present in the initial training set. To predict a single class, the model is queried with a sentence such as "a photo of a dog": the class is converted into a caption which can be paired with one or more images in the test set.

The idea behind CLIP is not new, but it became popular due to its strong generalization ability. CLIP zero-shot classification performed as state-of-the-art methods on popular image classification tasks, even when trained with a fewer number of labels; CLIP also gets a large margin accuracy improvement on out-of-distribution datasets.

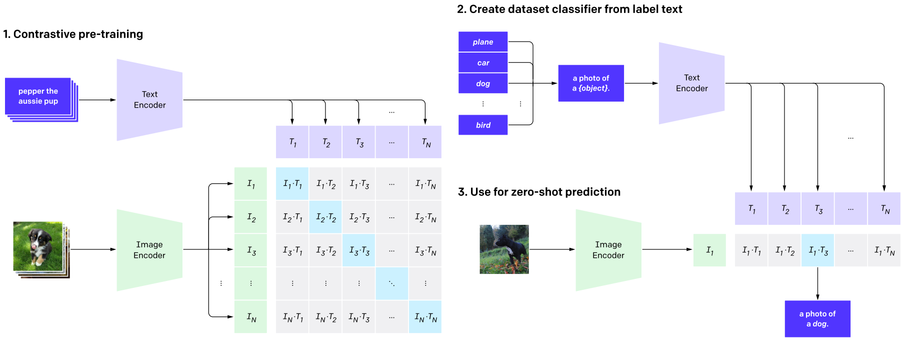
1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
$T_1$ $T_2$ $T_3$ ... $T_N$
$I_1$ $I_2$ $I_3$ ⋮ $I_N$
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$
2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
$I_1$
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
a photo of a dog.

**Fig. 4** The CLIP architecture [1], trained on a multiclass dataset for zero-shot classification.

CLIP authors suggest using their model as *pre-trainer* to transfer knowledge onto new image classification tasks, using ViT [5] or ResNet [6] as an image encoder, and a text transformer like GPT-3 [7]. The CLIP paper has also an extensive section about the previous work done on the topic, including ConVIRT [8] which did something similar on medical imaging datasets. Since then, many researchers have been working on a way to gain an advantage on visual tasks with the additional information given by text. We try to leverage some of the already available knowledge and apply it in the complex world of medical imaging.

# 4 Methods

For the purpose of this work, we implemented a simple architecture that might be developed further in the future to achieve better results. In order to learn a vector representation of brain MRI scans, we implement a simplified CLIP-like contrastive learning model with a text encoder and a 3D image encoder. After obtaining two features vectors through the respective encoders, we try to match each image-text pair by computing a similarity score and minimizing a contrastive loss. We have also trained a simple neural network to classify the resulting joint embedding and, to validate our results, we separately trained a image classifier and text classifier on the same data splits. However, these experiments are not reported because of inconclusive results.

## 4.1 Text encoder

To learn the text embedding, we use DistilBERT, a BERT-like pre-trained model from the HuggingFace library. DistilBERT was pretrained on the BookCorpus, a dataset consisting of 11,038 unpublished books and on the English Wikipedia. We don't fine tune the model, but train only a new additional layer that reduces the output of the model from 768 features to a common embedding size, to match the size of the image encoder output. More pre-trained text encoders are available, including ones fine-tuned specifically on medical data, but we use DistilBERT as a baseline for future work, especially for its reduced complexity. Before feeding a sample to DistilBERT, we have to tokenize it. Tokens are used to get a atomic representation of words from text, and they are later added to a dictionary leading to the original word. Words that do not appear in the original dictionary are referenced with a special token. The maximum number of tokens for the input of DistilBERT is 512, which means that any sequence over this limit is clipped to that size.

## 4.2 Image encoder

In order to learn a image embedding from batches of 3D image without developing a new architecture, we exploit the already available 3D UNet by Wolny et al. [9]. However, we only train the contracting path to reduce the input dimensionality, and discard the decoder path. The last convolution flattened output is connected to a fully connected layer. This encoder is the only one affected by training updates, as we leave the text encoder weights frozen. It is worth mentioning that this is the most difficult part of the network to train, while the text encoder is exploiting pretrained weights instead of random initialization. The image encoder building block were the ones implemented by Wolny et al. [9]. [2]

## 4.3 Contrastive learning

The essence of this learning method is given by the contrastive loss computation. We reduce the dimensionality of both the text and image to a vector of 256 values. In order to relate images and text, we try to minimize two losses, the text and the image loss. We get the cosine similarity of an image with its text by computing the dot product of the normalized image embedding and the normalized transposed text embedding. By doing this in batches, we obtain a image-text similarity matrix. If we transpose this matrix, we will instead get the similarity of a text with respect to its paired image. An image-text batch is perfectly matching when the two matrices multiplication equals the identity matrix. We compute the two cross entropy losses on the images-texts similarity matrix and on the texts-images similarity matrix with respect to an array of the same size, with entries being the position of the matching

**Listing 1** The forward function for the proposed method.

```
def forward(image, input_id_report, attention_mask_report, label):
        I_f = image_encoder(image)
        T_f = text_encoder(input_id_report, attention_mask_report)

        W_i = image_projection(I_f)
        W_t = text_projection(T_f)

        I_e, T_e = map(lambda t: F.normalize(t, p = 2, dim = -1), (W_i, W_t))
        sim = torch.einsum('i d, j d -> i j', I_e, T_e)
        sim *= temperature.exp()

        labels = torch.arange(I_f.size(0))
        I_loss = F.cross_entropy(sim, labels)
        T_loss = F.cross_entropy(sim.T, labels)

        loss = (I_loss + T_loss) / 2

        return loss
```

[2] https://github.com/wolny/pytorch-3dunet/blob/master/pytorch3dunet/unet3d/buildingblocks.py

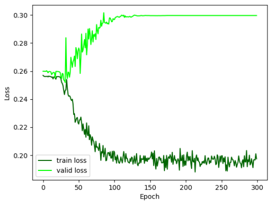

**Fig. 5** Training and validation loss for the proposed method, showing overfitting.

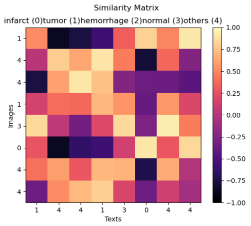

**Fig. 6** Example of a similarity matrix computed on one batch during training. The ideal matrix is the identity matrix, with values equal to one (bright) on the diagonal. However, when training on similar sample, we would also expect images with the same class to have higher similarity values.

sample, given by simply following the pair order in the batch. The two losses are averaged and used to compute the gradient. The implemented loss is described in Listing 1.

## 5 Experiments

Each model training parameters have been selected through trial and error. Further structured experiments, together with a larger dataset, are required to improve performance.

We train with the Adam optimizer with initial learning rate of 0.01, reducing on plateau with patience of 10 epochs on the validation loss by a factor of 0.5. The mini-batch size we tested was 8, which sounds reasonable with a training set of 211 images (0.04% of the training set). However, it is worth noticing that Radford et al. [1] used a mini-batch size of 32,768 on a training set of 400 million image-text pairs, being the 0.0008% of the total. We train our model for 300 epochs, however, from Figure 5 we can notice a strong overfitting at the beginning of training. We noticed an improve in performance when data augmentation was implemented, but we still believe that no hyperparameter tuning can overcome the lack of training data. Initially, the model was trained on only 89 images, of which half belonged to the class ”infarct” and the other half to healthy brains, in order to train a binary classifier. The only MRI sequence for each image was DWI. However, even if the complexity of the task was lower, the model assigned very small and indistinguishable similarity scores. Surprisingly, the model performed better with all three MRI sequences and different pathologies. This leads us to think that, in this context, a higher training set is preferred to the simplification of the initial hypothesis.

## 6 Results

The final model is able to return similarity scores of a given image or text input with respect to a $n$ number of test examples. One performance evaluation strategy would be that of finding the right prompt, based on the desired information; however we try a quantitative way to query the model. We test our method by querying the model with the report from each image-text pair of the test set ($n = 27$) onto the remaining

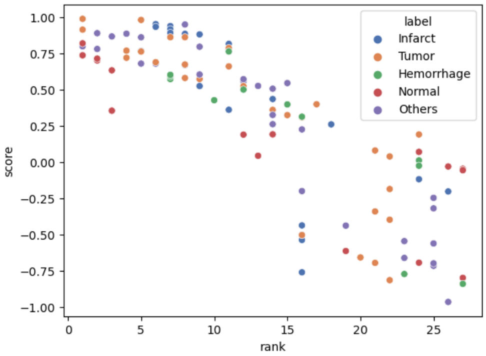

**Fig. 7** For each report in the test set ($n = 27$), we compute its similarity between each other sample ($n - 1$). We plot the similarity score for each pair-pair match only if they have the same ground truth class annotation. Ideally, a perfect match would have a rank and a similarity score close to 1 for the same class. If one sample has a high similarity and high rank, it means that the model was confident in the prediction. Even if most query results are wrong, we can still see that no sample had a big mismatch between similarity and ranked position, meaning that the model similarity are not diverging. This would happen in early stages of development, when every other sample was similar to each other.

samples ($n - 1$). Doing so, we can sort the obtained similarity scores in decreasing order, obtaining a rank with the highest position being the most similar image-text pair. If the system was working perfectly, we would match the queried report with itself in the top positions of the ranking. No image ranked itself in the first position, however we can see from Figure 7 that, plotting matches and grouping by class labels, the model tends to use all the available interval of [-1, 1] to describe the pairs. In the first versions of the model, the similarity scores were very similar to each other and close to zero, indicating that the model couldn't tell features apart. In Table 3, we can see which pair-pair match with the same available ground truth label was ranked in first position. The images of healthy brains are the easiest to match correctly. We also notice that two images matched each other reciprocally, indicating that the model has probably learned features that are common in both samples. In this case, both images show a tumor, in different but very close parts of the brain, the pituitary gland and the left prepontine cistern. This is not statistically relevant, but it is interesting behaviour. From Table 2 we can see that the learnt features don't rely completely on text syntax, but it seems that the model results are also based on all the available information: if the similarity score was based only on the presence of a word, it would be hard to pick up negations or irrelevant pathologies. Lastly, during data exploration and results evaluation, we noted that some of the ground truth labels were wrong or misleading. This could introduce noise in the model if this information was used in a classification task, or when evaluating performance.

# 7 Conclusion

We used a multi-modal input to train a contrastive learning model to learn a vector representation of MRI image-text pairs. Previous work relies on much bigger training

|  | Score | Report | Findings |
|---|---|---|---|
| 1 | 0.813 | Normal study of brain. No acute infarcts / haemorrhage / focal lesions. Vascular loop from right superior cerebellar artery is seen indenting the cisternal segment of right trigeminal nerve - suggested clinical correlation to r/o trigeminal neuralgia. | Normal |
| 2 | 0.770 | Few non-specific T2 FLAIR white matter hyperintensities are seen in bilateral frontal lobes. No significant abnormality in the brain, MR Angiogram and MR Venogram. | Normal |
| 3 | 0.767 | No significant abnormality is seen in the brain. | Normal |
| 4 | 0.760 | Grade I global cortical atrophy with scattered punctate white matter ischaemic changes. No acute infarcts / haemorrhage / focal lesions. Page 1 of 1 | Normal |
| 5 | 0.700 | Normal traumatic neuroparenchymal injury. No acute infarcts / haemorrhage / focal lesions. Subgaleal hematoma in the left frontal, parietal and temporal region with contusions in the left temporalis muscle. | Normal |

**Table 2** Highest ranking medical reports from matched image-text pair for the query ”No significant abnormality is seen in the brain.”

sets, or in the case of medical imaging, on single a 2D image, usually with only two classes of the form ”healthy” or ”not healthy” patient. We increase the complexity of the task by using all the available axial plane slices, and multiple MRI sequences. Our experiments showed inconsistent, but promising results. It would be possible to train a classifier on the learnt image or image-text embedding. Our experiments were not successful in producing meaningful results, as the model failed to converge during training. The reason could still be explained by the poor number of training data, or by the overall complexity of the task. Moreover, the ground truth available could be noisy, as it was inferred by a natural language processing algorithm, and not carefully validated by medical professionals. We believe that future work on this method can lead to two use cases: one would be that of zero-shot classification, which leads to more use cases in itself; the other is that of image retrieval, that can be used by health care professionals to compare cases, or by other machine learning engineers in increasing the diversity of their training set, or in helping with class imbalance by including more samples similar to the less frequent ones.

**Acknowledgments.** The data used to produce and test the method proposed in this study has been provided by *Cerebriu A/S* for research purpose only. Special thanks to Akshai Pai and Silvia Ingala for their supervision and support.

| **Image** | **Text (query)** | **Matched Image** | **Matched Text** | **Class** |
|---|---|---|---|---|
| image_236 | No significant abnormality is seen in the brain. | image_474 | No significant abnormality is seen in the brain. | Normal |
| image_576 | No significant abnormality is seen in the brain. | image_474 | No significant abnormality is seen in the brain. | Normal |
| **image_1069** | Sellar and suprasellar lesion indenting on inferior aspect of optic chiasma, likely pituitary adenoma. Suggest: Post contrast imaging. Mild small vessel ischemic changes. | **image_812** | Mildly enhancing, well-defined, oval, T2 heterogenous hypointense to isointense lesion at left prepontine cistern with a wide angle to the dura as described above, likely representing meningioma. Homogenous strongly enhancing, well-defined, rounded T1/T2 isointense lesion in left parafalcine region as described above. It may likely represent meningioma. Adv clinopathological correlation and review with relevant clinical details. | Tumor |
| image_387 | No significant abnormality is seen in the brain. | image_474 | No significant abnormality is seen in the brain. | Normal |
| **image_812** | Mildly enhancing, well-defined, oval, T2 heterogenous hypointense to isointense lesion at left prepontine cistern with a wide angle to the dura as described above, likely representing meningioma. Homogenous strongly enhancing, well-defined, rounded T1/T2 isointense lesion in left parafalcine region as described above. It may likely represent meningioma. Adv clinopathological correlation and review with relevant clinical details. | **image_1069** | Sellar and suprasellar lesion indenting on inferior aspect of optic chiasma, likely pituitary adenoma. Suggest: Post contrast imaging. Mild small vessel ischemic changes. | Tumor |
| image_989 | No significant abnormality in the brain. Hypoplastic intracranial right vertebral artery. Rest of MRA normal. Hypoplastic left transverse and sigmoid sinus. Rest of MRV normal. Kindly correlate clinically. | image_110 | Grade I global cortical atrophy with scattered punctate white matter ischaemic changes. No acute infarcts / hemorrhage / focal lesions. Page 1 of 1 | Others |
| image_411 | No significant abnormality is seen in the brain. Page 2 of 2 | image_576 | No significant abnormality is seen in the brain. | Normal |

**Table 3** For each report in the test set ($n = 27$), we compute its similarity between each other sample ($n - 1$). Here we show the medical reports for the images with the highest similarity ($rank = 1$) that also matched the query ground truth class and the matched pair class. In bold, the images that matched between one another. This is the desired behaviour, indicating that if two images in the test set display the same features, then they should be matched together when the corresponding reports are queried by the model. Three images with the "Normal (brain)" class all matched to the same image. In general, it seems that the "Normal" class was the easiest to learn and match.

# References

[1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. arXiv (2021). https://doi.org/10.48550/ARXIV.2103.00020 . https://arxiv.org/abs/2103.00020

[2] Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine, 106236 (2021) https://doi.org/10.1016/j.cmpb.2021.106236

[3] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv (2019). https://doi.org/10.48550/ARXIV.1910.01108 . https://arxiv.org/abs/1910.01108

[4] Cçiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. arXiv preprint arXiv:1606.06650 (2016)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv (2020). https://doi.org/10.48550/ARXIV.2010.11929 . https://arxiv.org/abs/2010.11929

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

[7] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. arXiv (2020). https://doi.org/10.48550/ARXIV.2005.14165 . https://arxiv.org/abs/2005.14165

[8] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. CoRR **abs/2010.00747** (2020) 2010.00747

[9] Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A.V., Louveaux, M., Wenzl, C., Strauss, S., Wilson-Sánchez, D., Lymbouridou, R., Steigleder, S.S., Pape, C., Bailoni, A., Duran-Nebreda, S., Bassel, G.W., Lohmann, J.U., Tsiantis, M., Hamprecht, F.A., Schneitz, K., Maizel, A., Kreshuk, A.: Accurate and versatile 3d segmentation of plant tissues at cellular resolution. eLife **9**, 57613 (2020) https://doi.org/10.7554/eLife.57613