

Research Internship Notes

my supervisor

me

email

my email

This is more of a log. Will probably show results of experiments , screenshots .. of the stuff that i would be working on [like a journal](#)

1. Agenda

1. DSRI project Id

2. Purpose of our article

This will keep changing, updated on :Mon Sep 30 10:24:29 AM CEST 2024

1. Make human readable report at radiologist clinics..
2. present current new and safer ways to get State of the art results for cheap.
 - Clinics/Hospitals instead of spending once should have yearly budget for local finetuning
 - as there are a lot of development on quantization and making the model smaller (budget balance with vol) while improving on the context window
 - present scaling law to calculate cost depending on the need
3. we will also present effective guarding techniques (phase after pre-training.. this is difficult..)
 - so that the report or Q/A does not give out horrible answers even if its right .. as it might be better if it came from a human
4. present a multimodality model for the ever developing need for adding new modalities. by developing a small **single** encoder-decoder model as opposed to many adapter models
 - see if we can improve masking techniques

3. Paligemma Transfer Hyperparameter Tuning

Mon Sep 30 10:24:29 AM CEST 2024

1. Did not see a difference with/without gradient accumulation
2. But changing target layers we train now 0.3% instead of just 0.1% of tot params

```
target_modules=["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"],  
  
# target_modules=["q_proj", "k_proj", "v_proj", "o_proj"],
```

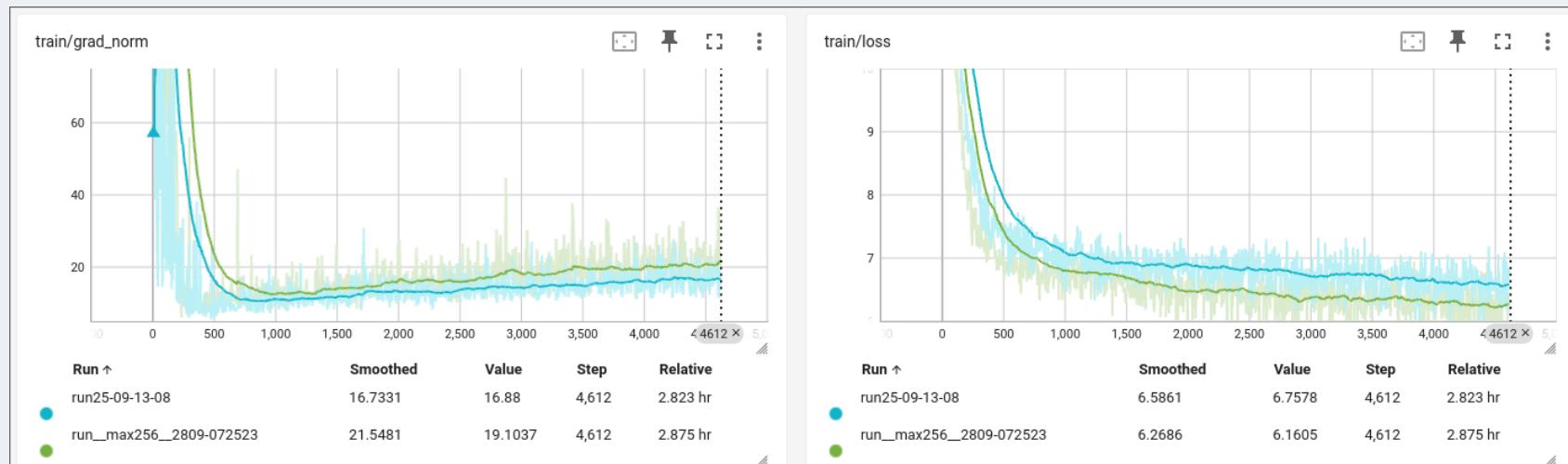


Figure 1: Green with more target modules

4. Ill conditioned Dataset ?..

mon Sep 30 11:33:01 AM CEST 2024

1. visual question answers are'nt supposed to give out long answers[even openended]?
 - Most of the model i read like TUM max token generation
 - With respect to the maximum text length accepted by CLIP, which is 76, **we trimmed any longer captions**, while zero-padding shorter ones. We refer to the resulting fine-tuned model as PubMedCLIP. PubMedCLIP was trained for 50 epochs
 - Pubmedclip

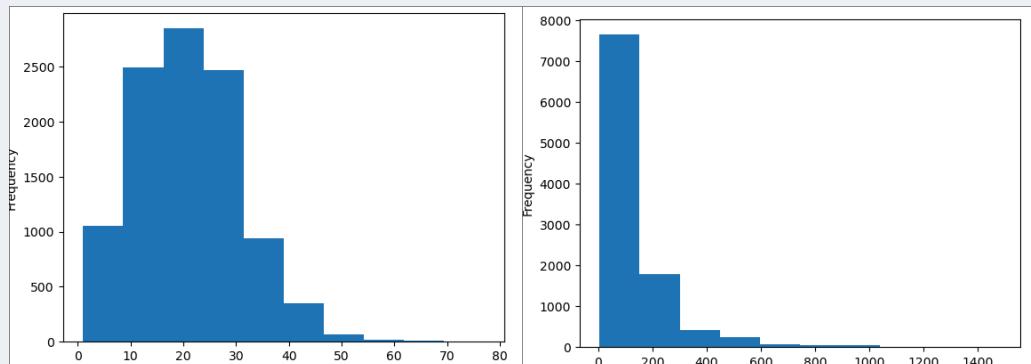


Figure 2: Left Generated by Llama [alignment] , Right Directly from captions [pre-training]

1. But good signs of getting better results than Pubmedclip
 - **we can accept a higher a context window size** : most of the results use 512. i am doing currently at 256

5. Epoch 1 results with context window 256

Monday Sep 30 11:33:01 AM CEST 2024

answer en what abnormalities does the patient have? The patient has brain swelling left atrium ventricle deformity left ventricle atrophy and the midline occipital a small ventricle.



{'U_id': 'MPX1014', 'image': 'MPX1014_synpic41745', 'caption': 'Axial T2 FLAIR demonstrates hyperintensity in the bilateral thalami, and periaqueductal grey matter. Post contrast imaging reveals enhancement in the periaqueductal grey matter, and bilateral mamillary bodies.', 'type': 'MRI', 'case_info': {'Title': 'Wernicke encephalopathy', 'History': '52 year old woman with a history of breast cancer, on chemotherapy. Poor oral intake and now stuporous.', 'Findings': '* T2 hyperintensity within the both thalamus* T2 hyperintensity of the periaqueductal grey matter\n* Enhancement of both mamillary bodies\n* Enhancement of the periaqueductal grey matter', 'Differential Diagnosis': '* bilateral thalamic glioma* metabolic/toxic processes (Wernicke encephalopathy, Leigh disease, Wilson disease)\n* infectious agents (Japanese and West Nile encephalitis, Creutzfeldt-Jakob disease)\n* vascular lesions(deep venous thrombosis, top of the basilar syndrome, infarction of the artery of Percheron)', 'Case Diagnosis': 'Wernicke encephalopathy', 'Diagnosis By': 'Pathology, laboratory, and imaging', 'Treatment & Follow Up': 'Patient was given thiamine, but expired', 'Discussion': 'Limited differential diagnosis - MR imaging is the modality of choice Differential diagnosis can be narrowed by the imaging appearance in combination with the presence or absence of other regions of involvement, & the patient's history. The differential includes primary neoplastic processes (bilateral thalamic glioma), metabolic/toxic processes (Wernicke encephalopathy, Leigh disease, Wilson disease), infectious agents (Japanese and West Nile encephalitis, Creutzfeldt-Jakob disease), & vascular lesions(deep venous thrombosis, top of the basilar syndrome, infarction of the artery of Percheron)', 'Topic_Info': {'Title': 'Wernicke encephalopathy', 'Disease Discussion': 'Severe neurologic disorder resulting from dietary vitamin B1 (thiamine) deficiency - frequently associated with chronic alcohol abuse. Thiamine is an essential coenzyme intermediate carbohydrate metabolism, and is also an osmotic gradient regulator. Deficiency may cause swelling of the intracellular space along with a local disruption of the blood-brain barrier. Wernicke encephalopathy is associated with chronic alcohol abuse, but can be caused by other disorders such as: gastropathy for obesity, psychogenic refusal of food, hyperemesis gravidarum, anorexia nervosa, voluntary food starvation, and parenteral therapy. The healthy body has a reserve of thiamine sufficient for up to 18 days, so any condition resulting in thiamine deficiency lasting 2 to 3 weeks may be an etiology\nIt is a medical emergency treated by intravenous administration of thiamine.Clinical triad of alteration in consciousness, ataxia, & ocular dysfunction\nImaging findings demonstrate symmetric T2 hyperintensity within the mamillary bodies, medial thalami , tectal plate, & periaqueductal grey.\nPost contrast enhancement may or may not be seen. Thiamine is osmotic gradient regulator, & deficiency may result in disruption of the blood brain barrier - leading to enhancement.Reduced diffusion on DWI may be seen. This may be related to ischemic-like changes occurring within the thalamus.', 'ACR Code': '1.5', 'Category': 'Toxic (see also Metabolic)', 'Keywords': 'Wernicke encephalopathyMamillary bodyThiamine', 'Reference': 'American Journal of Neuroradiology 29:164-169, January 2008'}}}

Figure 3: cherry picked good answer

answer en what abnormalities does the patient have? A 1 mm tumor of the left cecus and a and 1. A small renal cell nodule. the colon has a heterogeneous mass right cecum. cecus cecum., cecal mass. large left lower stomach. the colon. is abnormal and the cecum. There is also. the in the right cecum. cecum. The cecum. with a mass in bowel.,The colon.al has a right cecum.82 with the large cecum. cec. is small cecum. cec. and the large pelvis is. The cecum. cecus is to the colon. cec. is cec . cec. within the cec. cec. cec. cec. has. A is cec. in the cec. cec. is cec. cec. has. This is colon. cec. as the. The cec. cec. cec. as cec. cec. to a large bowel. as. of a cec. cec. (cec. right cec. cec. cec. for. cec. cec. in. cec. cec. cec. cec.) cec. cec. cec. cec. cec. cec.



{'U_id': 'MPX1009', 'image': 'MPX1009_synpic46295', 'caption': 'Bladder is prominent with mildly thickened wall. There is a small posterolateral diverticulum on the rightward aspect.', 'type': 'TAC', 'case_info': {'Title': 'Bladder Diverticulum', 'History': '73-year-old male with hematuria and numerous white blood cells found on UA', 'Exam': 'N/A', 'Findings': 'Bladder with thickened wall and diverticulum on the right. Diverticulum is mostly likely secondary to chronic outflow obstruction.\n\nProstate enlargement.', 'Differential Diagnosis': 'Bladder Diverticulum', 'Case Diagnosis': 'Bladder Diverticulum', 'Diagnosis By': 'N/A'}, 'topic_info': {'Title': 'Bladder Diverticulum', 'Disease Discussion': 'Bladder diverticula most often occur as a result of outlet obstruction. Occasionally, a congenital weakness in the bladder wall adjacent to the ureteral orifice results in a diverticulum. This is termed a "Hutch" diverticulum.\n\nIn children, outlet obstruction causing a diverticulum is rare and can be seen with urethral valves. In men, diverticula are associated with outlet obstruction from urethral stricture, prostatic hypertrophy, prostatic carcinoma etc. acquired diverticula are rare in women.\n\nDiverticula usually occur on the lateral bladder walls, rarely the dome. They are often multiple. Large diverticula often displace the bladder and or ureters. Diverticula can have wide or narrow necks. The wide necked variety empty urine readily. The narrow neck type are slow to empty and therefore are more likely to have urinary stasis.\n\nInfection, tumor and stone formation can occur as a result of urine stasis within a diverticulum. Tumor formation in a diverticulum is more likely to spread beyond the bladder because the diverticulum wall consists only of urothelium without muscle.\n\nBladder diverticula can be evaluated with excretory urography, ultrasound, CT and cystoscopy.\n\nRef:\nDunnick, R., McCallum, R., Sandler, C., Textbook of Uroradiology', 'ACR Code': '8.9', 'Category': 'Diverticulum'}}}

Figure 4: cherry picked bad answer

6. Scaling Law

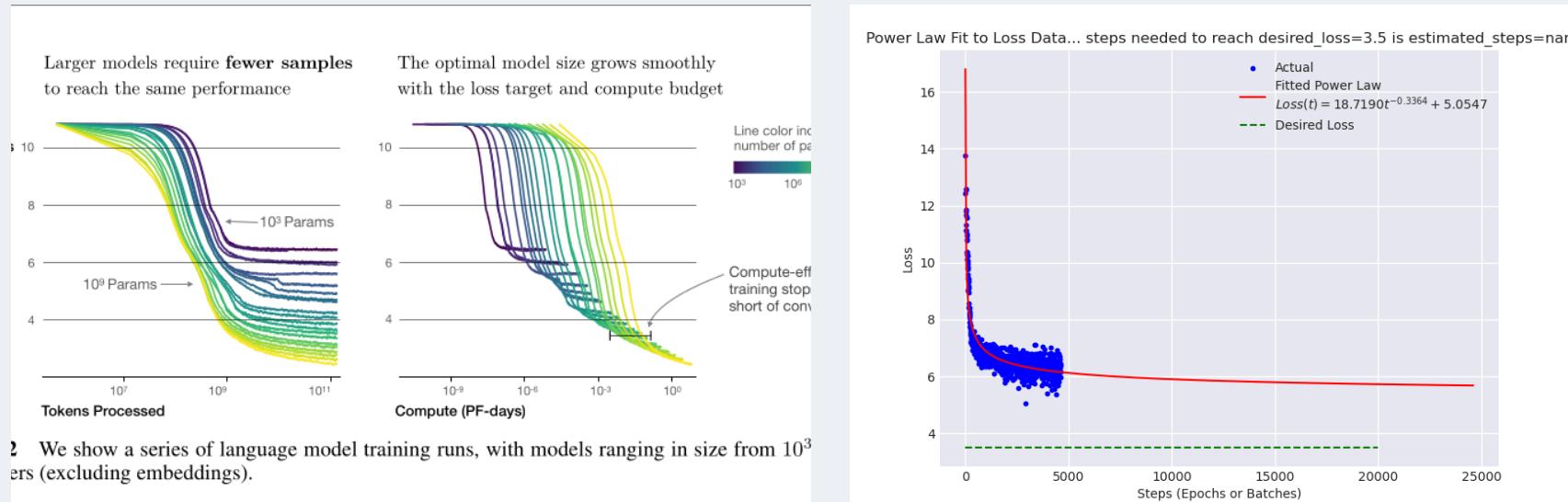


Figure 5: Scaling law

$$y = a(t^{-b}) + c$$

1. We will never get to the desired loss with our current hyperparameters/dataset
2. shows theoretical as 5.05 ...
3. we will do this again.. [possibly miscalculated]

draft

7. Done

<|end of notes|>

Second Meeting

- 26th Sep / Thursday
-

so we will go over :

- * Last Week
- * This Week
- * Quick Demo from yesterday's checkpoint [very likely that it might be funny]

Administrative stuff

- vpn [still 😊]

💡 **On Linux:** use `openconnect` to connect to the UM VPN. You can easily install it on Ubuntu and Debian distributions with `apt`:

```
sudo apt install openconnect  
sudo openconnect -u YOUR.USER --authgroup 01-Employees vpn.maastrichtuniversity.nl
```

🍏 **On MacOS and Windows:** download and install the **Maastricht University VPN** client available at vpn.maastrichtuniversity.nl

▼ ⚠️ If you are a **student** you will need to request access to the UM VPN first

- You can try to use the Athena Student Desktop at athenadesktop.maastrichtuniversity.nl, to access the VPN through a virtual desktop
- Or ask one of your teachers to request VPN access for you. You will need to send an email to the IT helpdesk of your department with the following information:
 - Email of the student who will get VPN
 - for which course (provide the course ID) or project does the student need the VPN
 - until which date the student will need the VPN.

:::

from previous meeting

1. who would use our model ? [grad students / radiologist / non medical]
1. for multimodality : I said segmentation.. and you asked why ? [and i didnt have an answer]

2. Research gaps [besides using different kind of models]

3. Areas of improvement [efficiency or usage]

1. [Radiologist and patients][hope its not generic.. this is still very informal]
-

Radiologists typically know which bone or anatomical region they will be examining when they suggest the patient get a CT scan or an MRI

[so they don't need image segmentation]

However, the patient might.

[better in clinics which have long latency]

- After the scan, the patient can receive a copy of the scan and our model while they wait to see their doctor.
- During this time, they can use our model to educate themselves on these basic, broad labels. This will help them prepare better, more informed questions for their doctor.

Potential questions the patient might ask:

- What am I looking at? [would allow us to do : semantic segmentation]
 - How does a healthy [xyz] look? [allows us to do: text to image but since they can look a bit cartoonish an even better option would be for the clinic to provide a healthy example of the [xyz] label for comparison, instead of them relying on googling image]
 - What condition might I have? [allows us to do VQA but we will also have to release a guarded model so that it doesn't say horrible stuff while they wait for the doctor .. or more intensive alignment after pretraining]
-

After consultation, we can have a human readable report generation:

- The scan image alongside its segmentation to ensure clarity in discussions.
 - A summary of the questions the patient asked, providing them with a clear record of the conversation.
 - and more... [this can probably be different for different clinics]
-

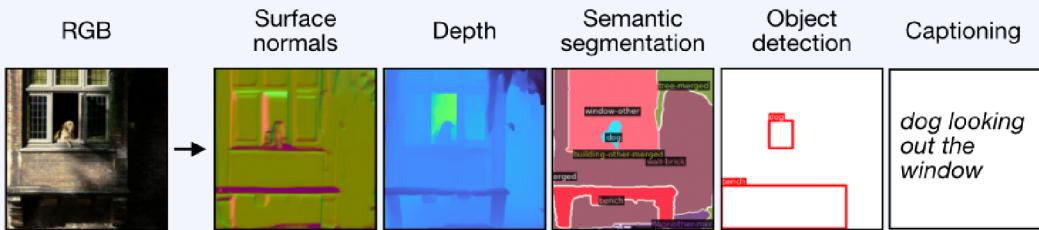
But since 4m can do many modality generation.

- we only need to give them 1 model → so they won't need as much resource.

- and in future if they need support for other modality → they can finetune themselves solving their privacy issues
- ... more maybe

A generalist vision model that can...

... perform a diverse set of vision tasks out of the box



For improvement :

- masking [there is a lot of masking]
 - unified masking but not fused.
 - even we can improve by percentage this could save a lot. [since the generic models are pretrained on Millions of images]
 - 4m did on CC12M

Last Week

- Released a VQA Dataset
- we had a lot of captions

```
// U_id is just patient id . each patient might have multiple images .
// but we are not doing that ...😊

// captions per image
{
  "Type": "CT",
  "U_id": "MPX1009",
  "Caption": "A CT scan of the brain showing a large, well-defined intracranial mass lesion with heterogeneous internal density, likely representing a glioma or similar neoplasm. The mass is causing significant displacement and compression of surrounding brain tissue. There is associated vasogenic edema and midline shift visible on the image."}
```

```
"image": "MPX1009_sympic46283",
"Description": {
    "ACR Codes": "8.-1",
    "Age": "73",
    "Caption": "The prostate is enlarged with several calcifications noted within. No dominant prostate mass is evident.",
    "Figure Part": null,           ----- This is almost always null
    "Modality": "CT - noncontrast",
    "Plane": "Coronal",
    "Sex": "male"
},
"Location": "Genitourinary",
"Location Category": "Reproductive and Urinary System"
},  

// -----  

// case captions [about the case the individual has][history][exams done][findings]
[literature]
{
    "Case": {
        "Title": "Bladder Diverticulum",
        "History": "73-year-old male with hematuria and numerous white blood cells found on UA",
        "Exam": "N/A",
        "Findings": "Bladder with thickened wall and diverticulum on the right. Diverticulum is mostly likely secondary to chronic outflow obstruction.\n\nProstate enlargement.",
        "Differential Diagnosis": "Bladder Diverticulum",
        "Case Diagnosis": "Bladder Diverticulum",
        "Diagnosis By": "N/A"
    },
    "Topic": {
        "Title": "Bladder Diverticulum",
        "Disease Discussion": "Bladder diverticula most often occur as a result of outlet obstruction. Occasionally, a congenital weakness in the bladder wall adjacent to the ureteral orifice results in a diverticulum. This is termed a \"Hutch\" diverticulum.\nIn children, outlet obstruction causing a diverticulum is rare and can be seen with urethral valves. In men, diverticula are associated with outlet obstruction from urethral stricture, prostatic hypertrophy, prostatic carcinoma etc. acquired diverticula are rare in women.\nDiverticula usually occur on the lateral bladder walls, rarely the dome. They are often multiple. Large diverticula often displace the bladder and or ureters. \nDiverticula can have wide or narrow necks. The wide necked variety empty urine readily. The narrow neck type are slow to empty and therefore are more likely to have urinary stasis.\nInfection, tumor and stone formation can occur as a result of urine stasis within a diverticulum. Tumor formation in a diverticulum is more likely to spread beyond the bladder because the diverticulum wall consists only of urothelium without muscle.\nBladder diverticula can be evaluated with excretory urography, ultrasound, CT and cystoscopy.\n\nRef:\nDunnick, R., McCallum, R., Sandler, C., Textbook of Uroradiology."
    }
}
```

```
        "ACR Code": "8.9",
        "Category": "Diverticulum"
    },
},
```

- Medpix also demos a vqa model [screenshot from the paper]

introducing sub-figure separation.

VQA-RAD [6] is a data set derived from MedPix®, and it collects a subset of radiological images, while providing Question-Answer (QA) pairs validated by domain experts.

- I was struggling to load in the layers of the model
- Layerstats [base model]:

```
adi@adi ~/code/um/sem3/res_internship/ours/transfer_pgMedpix (main*)>
python paligemma_layerstats.py
```

Total model size: 5.4454 GB

But this didnt even fit colab gpu [T4 Gpu 16 Gigs of HBM]

hugging face accelerate [mostly for inference]

Working with large models

Dispatch and offload

allows certain layers to be :

- cpu offload
- disk offload

so i thought this is great i can train in full resolution in fp32. if i fit 1 layer at a time albeit it would have been slow

But i couldnt

because we can't because we need 4 times the layer size to train the layer.

- 1 layer
- 1 gradients
- 2 optimizer states (adam also holds previous gradients)

- LORA training

LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

Edward Hu* Yelong Shen* Phillip Wallis Zeyuan Allen-Zhu
Yuanzhi Li Shean Wang Lu Wang Weizhu Chen
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency.

-- From Abstract

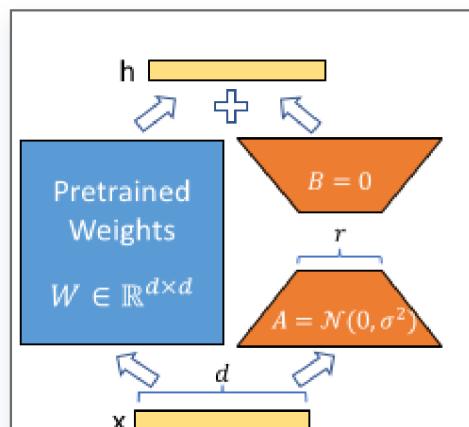


Figure 1: Our reparametrization. We only train A and B .

so now we only have to store 4 times of `dxr` and `rxr` which is smaller

than 4 times of `dxd`

as `r << d`

And here is how it compares to finetuning :

Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB _{base} (FT)*	125.0M	87.6	94.8	90.2	63.6	92.8	91.9	78.7	91.2	86.4
RoB _{base} (BitFit)*	0.1M	84.7	93.7	92.7	62.0	91.8	84.0	81.5	90.8	85.2
RoB _{base} (Adpt ^P)*	0.3M	87.1 _{±.0}	94.2 _{±.1}	88.5 _{±1.1}	60.8 _{±.4}	93.1 _{±.1}	90.2 _{±.0}	71.5 _{±2.7}	89.7 _{±.3}	84.4
RoB _{base} (Adpt ^D)*	0.9M	87.3 _{±.1}	94.7 _{±.3}	88.4 _{±.1}	62.6 _{±.9}	93.0 _{±.2}	90.6 _{±.0}	75.9 _{±2.2}	90.3 _{±.1}	85.4
RoB _{base} (LoRA)	0.3M	87.5 _{±.3}	95.1 _{±.2}	89.7 _{±.7}	63.4 _{±1.2}	93.3 _{±.3}	90.8 _{±.1}	86.6 _{±.7}	91.5 _{±.2}	87.2
RoB _{large} (FT)*	355.0M	90.2	96.4	90.9	68.0	94.7	92.2	86.6	92.4	88.9
RoB _{large} (LoRA)	0.8M	90.6 _{±.2}	96.2 _{±.5}	90.9 _{±1.2}	68.2 _{±1.9}	94.9 _{±.3}	91.6 _{±.1}	87.4 _{±2.5}	92.6 _{±.2}	89.0
RoB _{large} (Adpt ^P)†	3.0M	90.2 _{±.3}	96.1 _{±.3}	90.2 _{±.7}	68.3 _{±1.0}	94.8 _{±.2}	91.9 _{±.1}	83.8 _{±2.9}	92.1 _{±.7}	88.4
RoB _{large} (Adpt ^P)†	0.8M	90.5 _{±.3}	96.6 _{±.2}	89.7 _{±1.2}	67.8 _{±2.5}	94.8 _{±.3}	91.7 _{±.2}	80.1 _{±2.9}	91.9 _{±.4}	87.9
RoB _{large} (Adpt ^H)†	6.0M	89.9 _{±.5}	96.2 _{±.3}	88.7 _{±2.9}	66.5 _{±4.4}	94.7 _{±.2}	92.1 _{±.1}	83.4 _{±1.1}	91.0 _{±1.7}	87.8
RoB _{large} (Adpt ^H)†	0.8M	90.3 _{±.3}	96.3 _{±.5}	87.7 _{±1.7}	66.3 _{±2.0}	94.7 _{±.2}	91.5 _{±.1}	72.9 _{±2.9}	91.5 _{±.5}	86.4
RoB _{large} (LoRA)†	0.8M	90.6 _{±.2}	96.2 _{±.5}	90.2 _{±1.0}	68.2 _{±1.9}	94.8 _{±.3}	91.6 _{±.2}	85.2 _{±1.1}	92.3 _{±.5}	88.6
DeBERT _{XXL} (FT)*	1500.0M	91.8	97.2	92.0	72.0	96.0	92.7	93.9	92.9	91.1
DeBERT _{XXL} (LoRA)	4.7M	91.9 _{±.2}	96.9 _{±.2}	92.6 _{±.6}	72.4 _{±1.1}	96.0 _{±.1}	92.9 _{±.1}	94.9 _{±.4}	93.0 _{±.2}	91.3

Table 2: RoBERT_{base}, RoBERT_{large}, and DeBERT_{XXL} with different adaptation methods on the GLUE benchmark. We report the overall (matched and mismatched) accuracy for MNLI, Matthew’s correlation for CoLA, Pearson correlation for STS-B, and accuracy for other tasks. Higher is better for all metrics. * indicates numbers published in prior works. † indicates runs configured in a setup similar to Houldby et al. (2019) for a fair comparison.

```
adi@adi ~/code/um/sems3/res_internship/ours/transfer_pgMedpix <main*>
→ python lora_stats.py
```

Total quantized model size: 2.1095 GB

And i was able to fit this even on my gpu

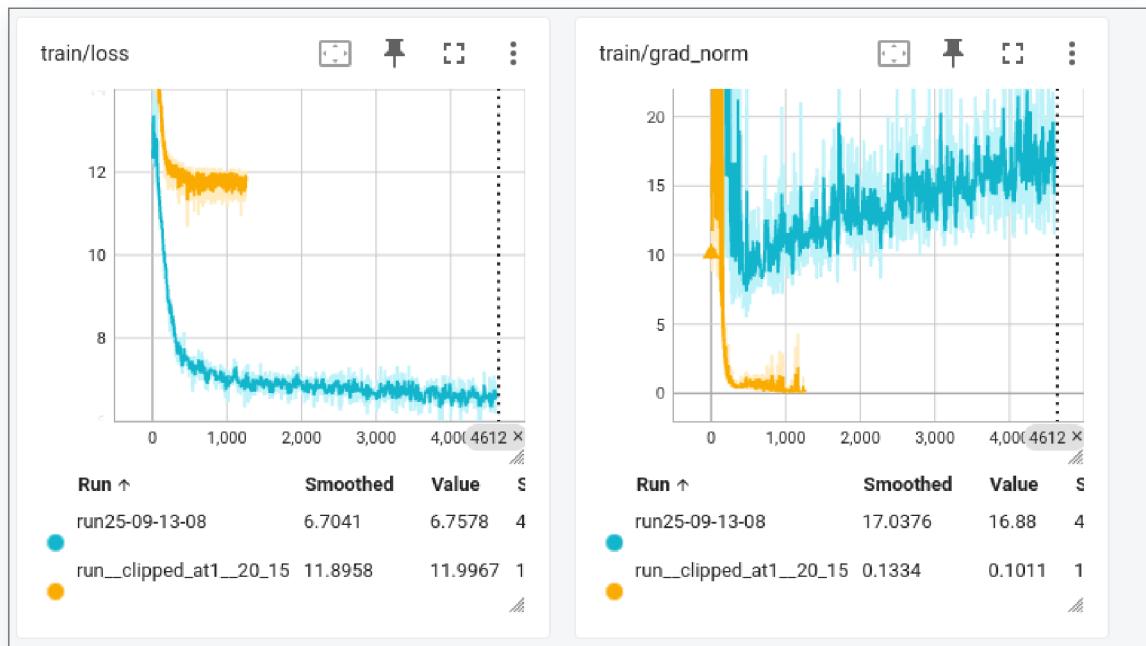
This week

- 1 epoch pretrain [experimentation with unreleased hyperparameters]

A Generalization of Full Fine-tuning. A more general form of fine-tuning allows the training of a subset of the pre-trained parameters. LoRA takes a step further and does not require the accumulated gradient update to weight matrices to have full-rank during adaptation. This means that when applying LoRA to all weight matrices and training all biases² we roughly recover the expressiveness of full fine-tuning by setting the LoRA rank r to the rank of the pre-trained weight matrices. In other words, as we increase the number of trainable parameters³ training LoRA roughly converges to training the original model, while adapter-based methods converges to an MLP and prefix-based methods to a model that cannot take long input sequences.

I missed a detail on gradient accumulation ?

- clipping [with no accumulation] at 1 killed all the gradients ...



Quick Demo

- i wrote a quick inference on how we will load in a checkpoint.
- hope cuda does not run out of memory.

Research Internship Notes

Chang

me

her email

my email

This is more of a log. Will probably show results of experiments , screenshots .. of the stuff that i would be working on [like a journal](#)

1. Checkpoints

Wed Oct 9 14:00 PM CEST 2024

- [✓] DSRI Access
- [] push code to github
 - ▶ submodules / standalone
 - dataset generation
 - finetuning
 - 4m
- [] Overleaf
 - ▶ [✓] created a project
 - ▶ [] have not made sections which would show the structure of our report
- [] 4m paper summary
 - ▶ not ready

2. Purpose of our article

This will keep changing, updated on :Mon Sep 30 10:24:29 AM CEST 2024

1. Make human readable report at radiologist clinics..
2. present current new and safer ways to get State of the art results for cheap.
 - Clinics/Hospitals instead of spending once should have yearly budget for local finetuning
 - as there are a lot of development on quantization and making the model smaller (budget balance with vol) while improving on the context window
 - present scaling law to calculate cost depending on the need
3. we will also present effective guarding techniques (phase after pre-training.. this is difficult..)
 - so that the report or Q/A does not give out horrible answers even if its right .. as it might be better if it came from a human
4. present a multi modality model for the ever developing need for adding new modalities. by developing a small **single** encoder-decoder model as opposed to many adapter models
 - see if we can improve masking techniques

3. Quick Recap

Wed Oct 9 14:00 PM CEST 2024

1. we changed target layers to train now 0.3% instead of just 0.1% of tot params

```
1 # current target layers for lora
2 target_modules=["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"],
3 # previous
4 # target_modules=["q_proj", "k_proj", "v_proj", "o_proj"],
```

python

2. we saw it was saturating with previous target layers even with 2 epochs

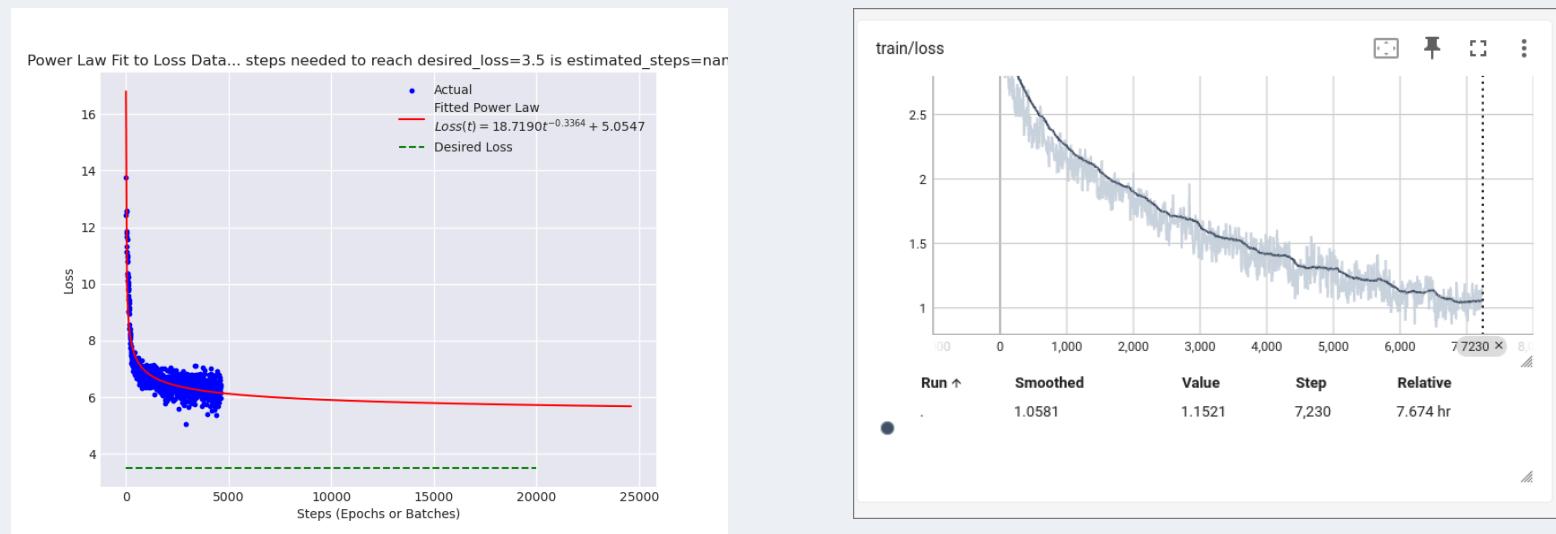


Figure 1: Left: Saturating after 2 epochs ; Right : Trained For 10 epochs

3. I let it run for 50 epochs on DSRI

4. Cherry Picked Results Epoch 10 checkpoint

https://huggingface.co/datasets/adishoucya/results__fullrun_0310-134147__infer

(1) input image (2) patient id (3) generated ans (4) true label

	MPX1298	answer What are some diagnostic modalities used for thoracic aortic aneurysms? CT, Echocardiography, and MRI.	Aortography is considered the gold standard, but contrast-enhanced CT, MRI, MRA, trans-thoracic echocardiography, and trans-esophageal echocardiography can also be used to diagnose these aneurysms.
	MPX2142	answer What is the likely diagnosis for the centrally necrotic circumscribed extraaxial mass located midline over the planum sphenoidale? The likely diagnosis is a meningioma, given its location and characteristics described in the case information.	The likely diagnosis is a meningioma, given the location and characteristics of the mass.
	MPX2142	answer What is a "dural tail" sign and how is it associated with meningiomas on MRI imaging? A "dural tail" sign is a radiologic finding where the meningioma extends beyond the limits of the dura, which is a normal finding on MRI imaging.	A "dural tail" sign refers to the appearance of tissue tapering into the dura, which is often seen in meningiomas on MRI imaging, particularly on T1-weighted images.
	MPX2519	answer How does a CT scan help diagnose massive pulmonary embolism? A CT scan can show areas of blockage in the pulmonary arteries, which can help confirm the diagnosis.	A CT scan can help diagnose massive pulmonary embolism by visualizing the blockage or clot in the pulmonary arteries, which may not be visible on plain radiographs due to its small size or location.
	MPX2519	answer What would be the likely outcome if a massive pulmonary embolism goes untreated? Death.	If a massive pulmonary embolism goes untreated, it can lead to severe complications such as cardiac arrest, respiratory failure, and even death due to the blockage of blood flow to the lungs and the resulting acute right ventricular heart failure.

5. Results

- google/paligemma-3b-mix-224

1. 10 Epochs

- Model Card https://huggingface.co/adishourya/results__fullrun__0310-134147
- Evaluation Results: https://huggingface.co/datasets/adishourya/results__fullrun__0310-134147__infer

2. 20 Epochs [Does worse than 10]

- Model Card https://huggingface.co/adishourya/results__fullrun__0710-111627
- Evaluation https://huggingface.co/datasets/adishourya/results__fullrun__0710-111627__infer__

3. 50 Epochs [Does worse than 20]

- Model Card https://huggingface.co/adishourya/results__fullrun__0710-151659
- Eval : Needless ..

horrible graph incoming!!... next page

6. horrible overfitting

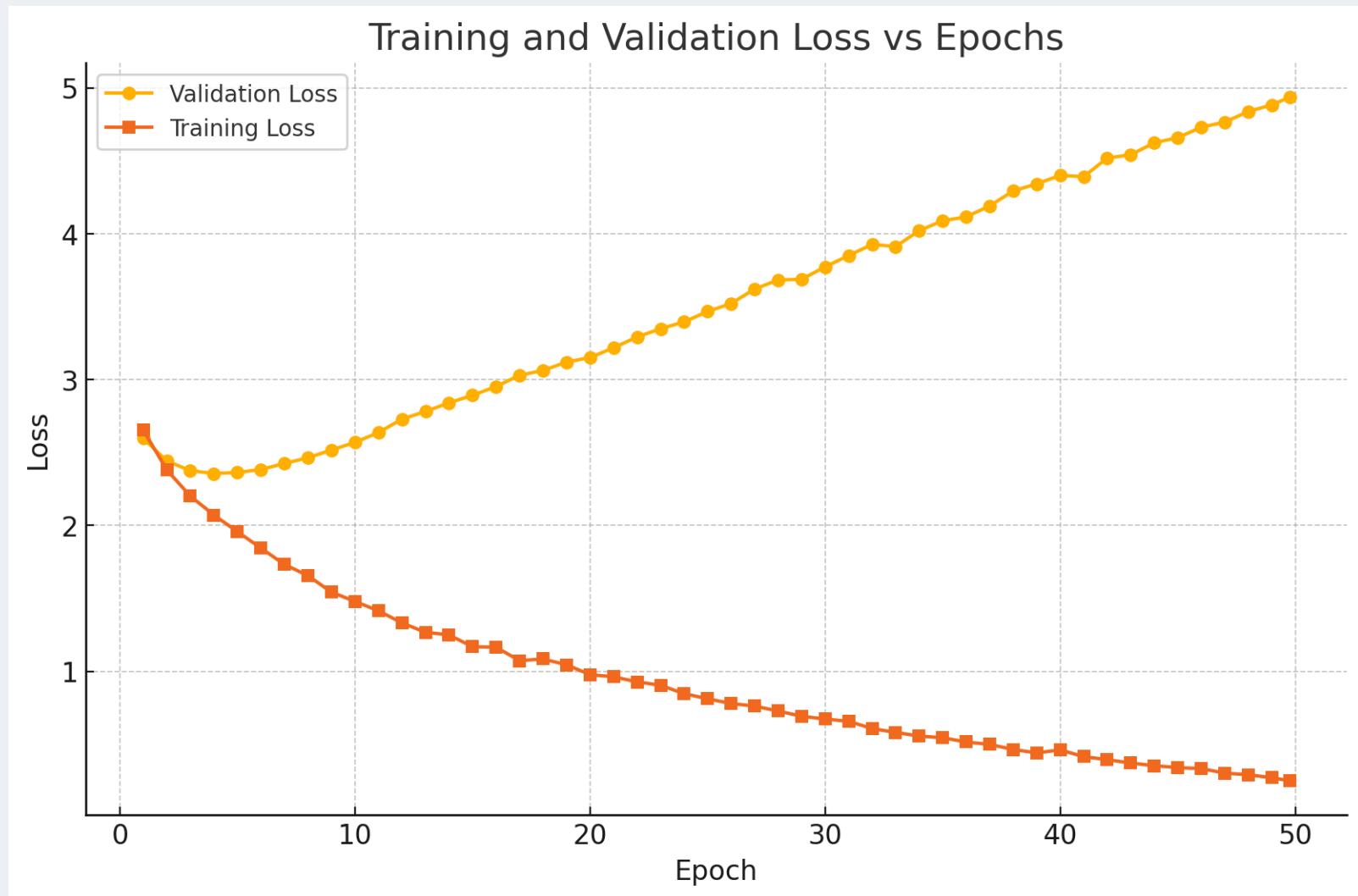


Figure 2: Overfitting Results .. most of my quickruns were on 2 epochs

7. To fix training ; Condition the dataset Better

1. current strategy : split on patients [650 -> 600/25/25]
2. Increase the quality of the visual question answering dataset. so that we could also formally present the dataset
 - Validation and Test Dataset could have conditions that might not even be present in training dataset
 - Remove questions that ask 3d information [page 16 MPX 2033 ; 3rd]
 - In general avoid answers with dimensions ?
 - Remove history questions [page 16 ; MPX 1879 ; 2nd]
 - More Questions on the literature [beginner level]
3. But we have time to iron out this list till the end of my project
4. I hope we can assume that it will get a lot better with current scale but better dataset
 - so I think.. no such technical challenges are left.
 - Blurry ? 448 just to test [will take 4-5 hours to run]

8. PHI3

1. Found an open source Visual Model .. for fair comparison <https://huggingface.co/microsoft/Phi-3-vision-128k-instruct>
2. People are using edit distance to compare llm style models
 - <https://www.sciencedirect.com/topics/computer-science/levenshtein-distance>

9. for the next meeting

1. if i progress more on finetuning
2. Masking section of 4m

<|END of Notes|>

Research Internship Notes

Chang

her email

me

my email

This is more of a log. Will probably show results of experiments , screenshots .. of the stuff that i
would be working on like a journal

1. Checkpoints

Thurs Oct 24 09:30 PM CEST 2024

- [✓] push code to github -> github.com/adishouya
 - ▶ submodules [✓] / standalone
 - dataset generation
 - finetuning
 - 4m
- [✓] 448 paligemma finetune (not good)
 - ▶ [] LLava-med model which was trained on PMC-MED 15M but license does not allow commercial use (does not follow our purpose of the report)
 - But we can use it just to see if our dataset is feasible . and maybe we can list it in the paper as a benchmark , and make a column of license in the table
 - ▶ [✓] collected compute metrics to run
- PMC Dataset :

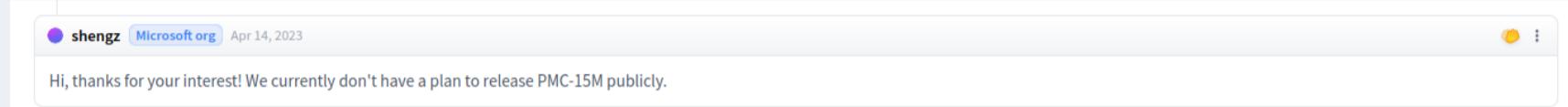


Figure 1: Like Conceptual Caption (CC12M google) dataset for medical images

- [✓] 4m Paper Summary (with snippets of code)

2. Purpose of our article

This will keep changing, updated on :Mon Sep 30 10:24:29 AM CEST 2024

1. Make human readable report at radiologist clinics..
2. present current new and safer ways to get State of the art results for cheap.
 - Clinics/Hospitals instead of spending once should have yearly budget for local finetuning
 - as there are a lot of development on quantization and making the model smaller (budget balance with vol) while improving on the context window
 - present scaling law to calculate cost depending on the need
3. we will also present effective guarding techniques (phase after pre-training.. this is difficult..)
 - so that the report or Q/A does not give out horrible answers even if its right .. as it might be better if it came from a human
4. present a multi modality model for the ever developing need for adding new modalities. by developing a small **single** encoder-decoder model as opposed to many adapter models
 - see if we can improve masking techniques

3. How others do it

- There are lots of ways to do it so i only focused on some of the common ones

As part of the Llama 3 development process we also develop multimodal extensions to the models, enabling image recognition, video recognition, and speech understanding capabilities. These models are still under active development and not yet ready for release. In addition to our language modeling results, the paper presents results of our initial experiments with those multimodal models.

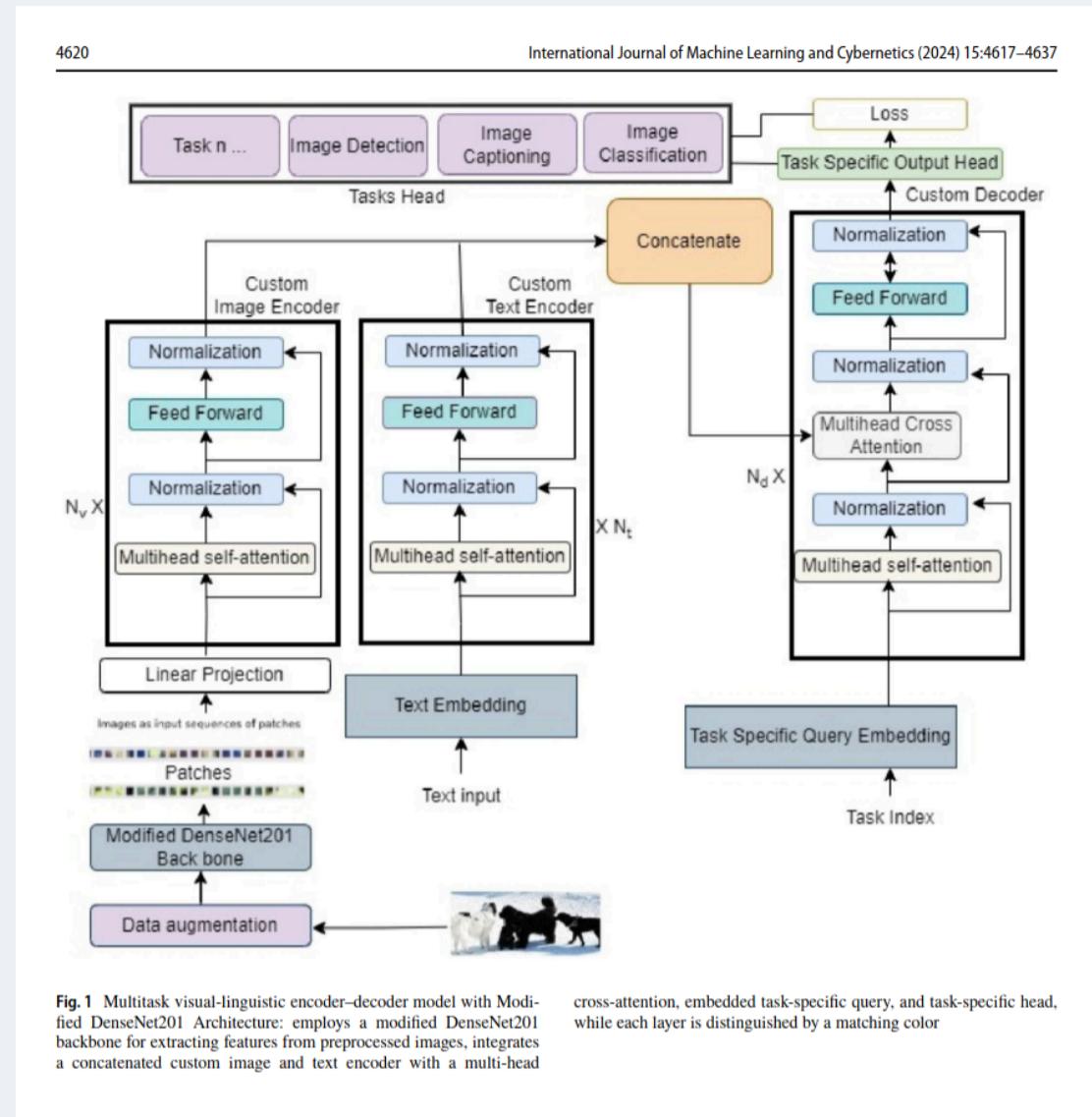


Fig.1 Multitask visual-linguistic encoder-decoder model with Modified DenseNet201 Architecture: employs a modified DenseNet201 backbone for extracting features from preprocessed images, integrates a concatenated custom image and text encoder with a multi-head cross-attention, embedded task-specific query, and task-specific head, while each layer is distinguished by a matching color

Figure 2: Unified deep learning model for multitask representation and transfer learning: image classification, object detection, and image captioning

1. There is 1 encoder per modality . And no interaction between encoder blocks [self attention throughout]
2. The key value pairs come after the concatenation of the output of the encoder blocks
3. for adding new modalities we would have to add a new encoder ;then the previous encoders could be frozen . and only the new encoder and decoder would be trained
4. So there are as many embedding space as modalities ; and there is a task specific mlp head • classification task : num_classes , captioning task : token max_generation

Table 5 Image captioning training on Flickr8k and Flickr30k

Dataset	Loss	BLEU-4	RougeL	CIDEr
Flickr8k	0.55	41.64	0.84	74.5
Flickr30k	0.52	33.42	0.82	68.9

Figure 3: Evaluation Metrics used for seq2seq task

3.1. VILBERT Encoders only (Bidirectional Transformer)

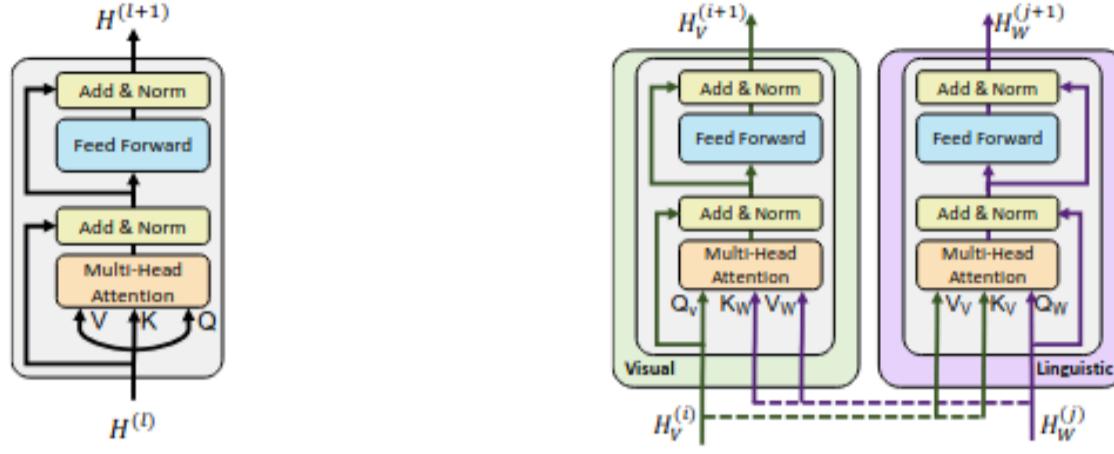


Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

Figure 4: VILBERT (Multiple)Encoder only

1 v is for visual
2 w is for words → linguistic

1. co-attention : (more like relative encoder decoder blocks) . The key value pair for linguistic encoder (here relatively decoder) comes from visual encoder (so purple block is the decoder for the green encoder block) and vice versa for the visual block
2. co-attention is very similar to bert in being bidirectional [This is a quick comparison b/n bidirectional and left-right transformers]

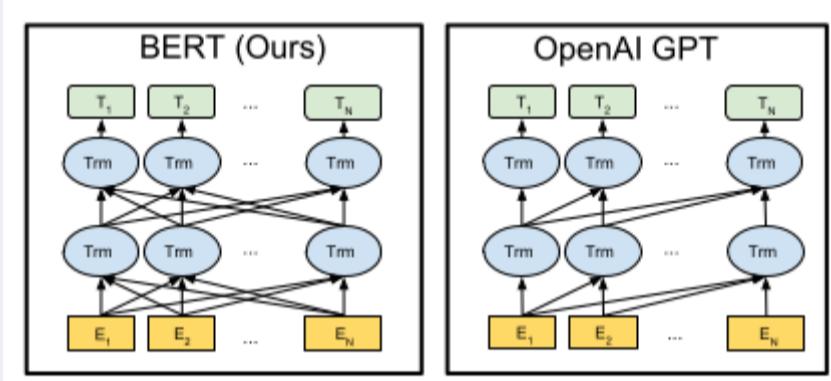


Figure 5: [Bert]Bert : Bidirectional ; GPT : Left to Right Transformer

- Straight line shows self attention and angled lines pointing inwards show the direction of passing kv pairs (cross attention in that node)
- [BERT] all nodes have cross attention
- [GPT] so the encoders to the right will have cross attention

Vision-and-Language. While we address many vision-and-language tasks in Sec. 3.2, we do miss some families of tasks including visually grounded dialog [4, 45], embodied tasks like question answering [7] and instruction following [8], and text generation tasks like image and video captioning [5]. These tasks may also benefit from a self-supervised approach similar to what we have presented. There are open questions on how to incorporate long sequences of images and text found in dialog, embodied tasks, and video processing. Further, it is unclear how to effectively decode output text from our bidirectional model as existing greedy decoders like beam-search do not apply.

Figure 6: [VILBERT] Section 5

4. 4m [Massively Multimodal Masked Modelling - Single Unified Transformer [left to right]]

- No Result table with evaluation metrics for captioning tasks in their paper ...

1. need for multimodality in Medical AI :

- Unlike NLP, where language modeling on raw text has led to multitask capabilities, training on only RGB images with a single objective has not exhibited the same behavior for vision. Therefore, it is deemed important to incorporate multiple modalities and tasks in training. It has been indeed suggested by psychophysical studies that multimodality is one key driver behind the development of biological intelligence

- Introduction

2. To train a single encoder decoder transformer it will have to be different than other multimodal transformers as we can only afford 1 representation space and 1 (task free) mlp head

- Tokenizing modalities: We abstract away modality-specific intricacies by mapping all modalities into sequences or sets of discrete tokens, whether they are images, text, sparse data, or neural network feature maps. This allows every possible mapping between modalities to be seen as predicting one sequence or set of tokens from another.\ Tokenizing all modalities into a unified representation space allows us to train a single Transformer encoder-decoder to map between different modalities through (parallel or serialized autoregressive) token prediction

- Section 2

- we are allowed to use specialized tokenizer but the objective needs to learn a single embedding space like VILBERT did.

- This tokenization approach enhances compatibility, scalability, and sharing by removing the need for task-specific encoders and heads, allowing the Transformer to be compatible with all modalities and maintain full parameter-sharing.

- Introduction

- so we dont need task specific heads at the end like we did before

3. So the inputs and outputs would also have to be **masked** and prepared slightly diffently

- Also, although 4M operates on a large set of modalities, it can train in a highly efficient manner through the use of input and target masking. This involves randomly selecting a small subset of tokens from all modalities as inputs to the model, while another small subset of the remaining tokens is treated as targets.

- Introduction

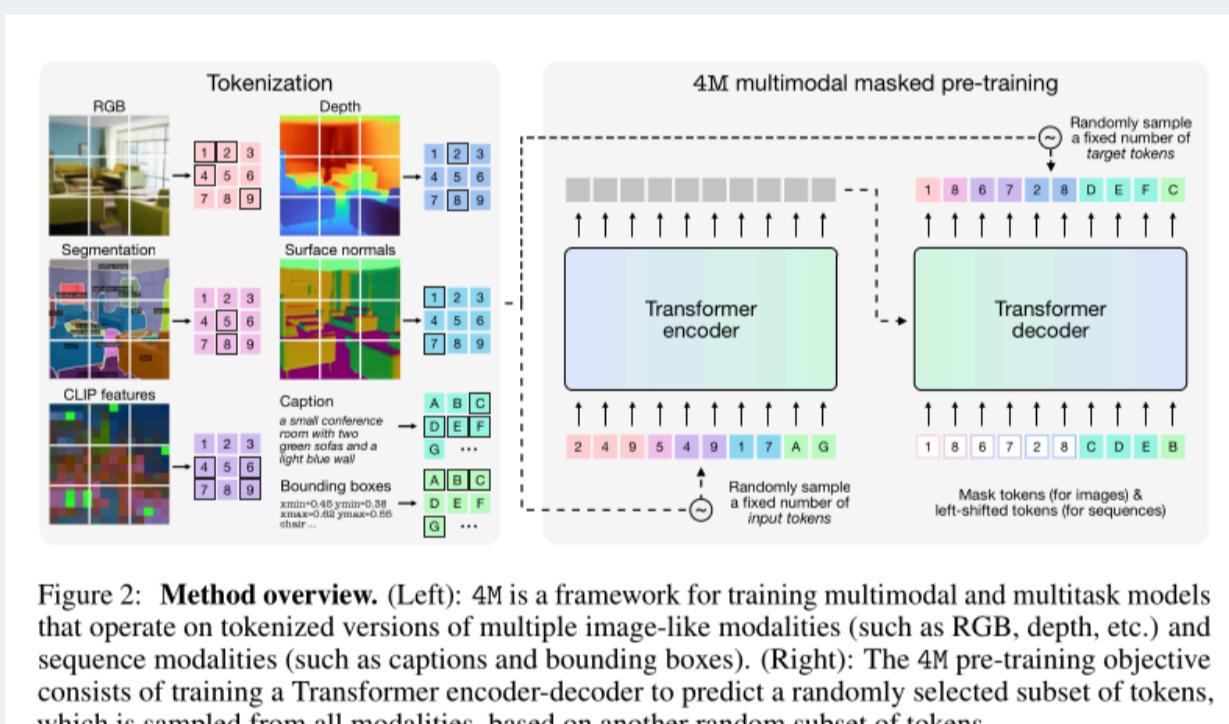


Figure 2: **Method overview.** (Left): 4M is a framework for training multimodal and multitask models that operate on tokenized versions of multiple image-like modalities (such as RGB, depth, etc.) and sequence modalities (such as captions and bounding boxes). (Right): The 4M pre-training objective consists of training a Transformer encoder-decoder to predict a randomly selected subset of tokens, which is sampled from all modalities, based on another random subset of tokens.

Figure 7: [4m]Different dataset preparation for Single Encoder Decoder System

1. Summary of what we see in the image:

- all modalities gets indexed by modality specific definition ; integers for iamge based, characters for sequence based modalities **indexed not tokenized!**
- sample of indices are drawn to generate an example ; input and target sequences (in black borders)
 - some of the sampled indexes are used as the input to the encoder
 - the remaining are used to make the target sequence (output of the decoder)
 - left to right transformers have generative capabilities.
 - for image based inputs : almost like autoencoder (noise fed in input)
 - dense modalities are masked from sequence inputs . so no self attention across modalities in decoder block. But they are free to tend to all key value pairs from the encoder (i.e all modalities)
 - for sequence based : right shifted output
 - here 3 from caption modality as input ; so it should be 3 from caption in the output [C D E] -> [D E F]
 - similarly for bounding box ; only 1 right shifted output [B] -> [C]
 - so there is a budget that controls number of tokens based on modalities

1. joint representational space from indices to vector , and backbone like modified densenet for image features

- Multimodal encoder. The encoder is a standard Transformer encoder but features modality-specific learnable input embedding layers to map token indices to vectors. To each token of a specific modality, we add a learnable modality embedding and either 1D (for sequences) or 2D (for dense modalities) sine-cosine positional embeddings. To facilitate transfer learning, the encoder is additionally designed to accept RGB pixels using a learnable patch-wise linear projection, enabling it to double as a Vision Transformer backbone.

- section2.2

2. job of decoder

- Multimodal decoder. The decoder handles tokens from both dense image-like and sequence-like modalities, with each type requiring a different approach. However, two aspects are common to all tokens: First, they can all freely attend to any encoder tokens in the cross-attention layers, ensuring full access to the encoded information. Second, we employ attention masks to separate decoder tokens of different modalities. This ensures that the decoder produces consistent outputs for each specific modality, irrespective of what other outputs are being generated simultaneously. For dense image-like modalities, the decoder input consists of mask tokens along with modality and positional information. The decoder's role is to predict this masked content. For sequence-like modalities, the input to the decoder comprises modality, positional, and content information. The decoder is tasked to predict the next token in the sequence. To ensure that each token is only influenced by preceding tokens (and not by any future tokens), we apply a causal mask to the self-attention, as is standard in autoregressive models. Since all target tasks consist of discrete tokens, we can use the cross-entropy loss for all of them, which we found removes the need for task-specific loss balancing and improves training stability.

- section2.2

3. These are some of the class documentations that makes it more clear and give better details

- Noise fed into image based inputs

```
1 def image_mask(self, tensor: torch.Tensor, num_tokens: int, input_budget: int, target_budget: int):  
2     """Applies input and target masking to an image tensor  
3  
4     Args:  
5         tensor: Image tensor  
6         num_tokens: Number of tokens in the tensor  
7         input_budget: Token budget for the input  
8         target_budget: Token budget for the target  
9  
10    Returns:  
11        Dictionary containing the masked image tensor, the input mask, the target mask, and the decoder attention mask  
12    """  
13    noise = torch.rand(num_tokens)  
14    ids_shuffle = torch.argsort(noise, dim=0)  
15    ...  
16  
17    return {"tensor": tensor, "input_mask": input_mask, "target_mask": target_mask, "decoder_attention_mask": decoder_attention_mask}  
18
```

- using all modalities to generate dataset

```
1 def build_fm_pretraining_dataset(  
2     data_path, all_domains, modality_info, modality_transforms,  
3     image_augmenter, text_tokenizer,  
4     input_tokens_range, target_tokens_range,  
5     sampling_weights=None):  
6     """Builds the FourM pre-training dataset based on the given arguments.  
7     This function should mainly used for smaller datasets (e.g. validation sets),  
8     while large training sets should be loaded with build_wds_fm_pretraining_dataloader in webdataset format.  
9  
10    Returns:  
11        FourM pre-training dataset as a PyTorch Dataset.  
12    """
```

- And then they create the mixture dataset

```
1 class MixtureDataset(IterableDataset):  
2     def __init__(self, data_iters, weights, modality_info):  
3         self.orig_data_iters = data_iters  
4         self.data_iters = [iter(data_iter) for data_iter in data_iters] # Create initial iterators  
5         self.sampling_probs = np.array(weights) / sum(weights)  
6         self.modality_info = modality_info
```

- From fourm/data/unified_dataset.py

1. This needs massive masking operation to create the dataset (800 lines of code)

- Decoupling the number of input and target tokens from the number of modalities prevents the computational cost from rapidly escalating with increasing modalities, allowing for a scalable training objective.

- Introduction

2. Sampling operations

```
1 def sample_cosine(min_val: float = 0, max_val: float =1) → float:  
2     """Sample a value from a cosine distribution between min_val and max_val"""  
3  
4     def sample_uniform(min_val: float = 0, max_val: float =1) → float:  
5         """Sample a value from a uniform distribution between min_val and max_val"""  
6  
7     class UnifiedMasking(object):  
8         def __init__(self,  
9             modality_info: Dict,  
10            text_tokenizer: Optional[Tokenizer],
```

```

11     input_tokens_range: Union[int, Tuple[int, int]],
12     target_tokens_range: Optional[Union[int, Tuple[int, int]]],
13     max_tries: int = 100,
14     sampling_weights: Optional[List[float]] = None,):
15     """Performs masking on a dict of modalities (both image based and sequence based modalities)
16
17     Args:
18         modality_info: Dict with the modalities and their corresponding information
19         text_tokenizer: Tokenizer to use for text modalities
20         input_tokens_range: Range of number of tokens to mask in the input
21         target_tokens_range: Range of number of tokens to mask in the target
22         max_tries: Maximum number of tries to find a valid token budgets
23         sampling_weights: Sampling weights for the mixture of Dirichlet distributions
24     """

```

- Pseudo labeled multimodal training dataset. Training 4M models requires a large-scale and aligned multimodal/multitask dataset that contains all the above modalities/tasks and is sufficiently diverse.

- Section 2

- but none of the sampling operations are fused and `torch.compile` does not fuse across these sampling operations . People usually write `[triton]` or `cuda` code with `torch.load_inline` to make it more efficient.
 - we could do that and discuss whether it improves the cost of sampling.

- a lot of papers write some pseudo code just to improve understanding but 4m paper doesn't . This is an example from siglip in numpy

```

1 # image_encoder - ResNet or Vision Transformer
2 # text_encoder - CBOW or Text Transformer
3 # I[n, h, w, c] - minibatch of aligned images
4 # T[n, 1]
5 # minibatch of aligned texts
6 # W_i[d_i, d_e] - learned proj of image to embed
7 # W_t[d_t, d_e] - learned proj of text to embed
8 # t
9 # learned temperature parameter
10 # extract feature representations of each modality
11 I_f = image_encoder(I) #[n, d_i]
12 T_f = text_encoder(T) #[n, d_f]
13 # joint multimodal embedding [n, d_e]
14 I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
15 T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
16 # scaled pairwise cosine similarities [n, n]
17 logits = np.dot(I_e, T_e.T) * np.exp(t)
18 # symmetric loss function
19 labels = np.arange(n)
20 loss_i = cross_entropy_loss(logits, labels, axis=0)
21 loss_t = cross_entropy_loss(logits, labels, axis=1)
22 loss
23 = (loss_i + loss_t)/2
24 # !Figure 3. Numpy-like pseudocode for the core of an implementation of clip

```

4.1. More Method Description

<|END of Notes|>

Research Internship Notes

my supervisor

email

me

my email

This is more of a log. Will probably show results of experiments , screenshots .. of the stuff that i would be working on
[like a journal](#)

1. Agenda

1. DSRI project Id

2. Purpose of our article

This will keep changing, updated on :Mon Sep 30 10:24:29 AM CEST 2024

1. Make human readable report at radiologist clinics..
2. present current new and safer ways to get State of the art results for cheap.
 - Clinics/Hospitals instead of spending once should have yearly budget for local finetuning
 - as there are a lot of development on quantization and making the model smaller (budget balance with vol) while improving on the context window
 - present scaling law to calculate cost depending on the need
3. we will also present effective guarding techniques (phase after pre-training.. this is difficult..)
 - so that the report or Q/A does not give out horrible answers even if its right .. as it might be better if it came from a human
4. present a multimodality model for the ever developing need for adding new modalities. by developing a small **single** encoder-decoder model as opposed to many adapter models
 - see if we can improve masking techniques

3. Paligemma Transfer Hyperparameter Tuning

Mon Sep 30 10:24:29 AM CEST 2024

1. Did not see a difference with/without gradient accumulation
2. But changing target layers we train now 0.3% instead of just 0.1% of tot params

```
target_modules=["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"],  
  
# target_modules=["q_proj", "k_proj", "v_proj", "o_proj"],
```

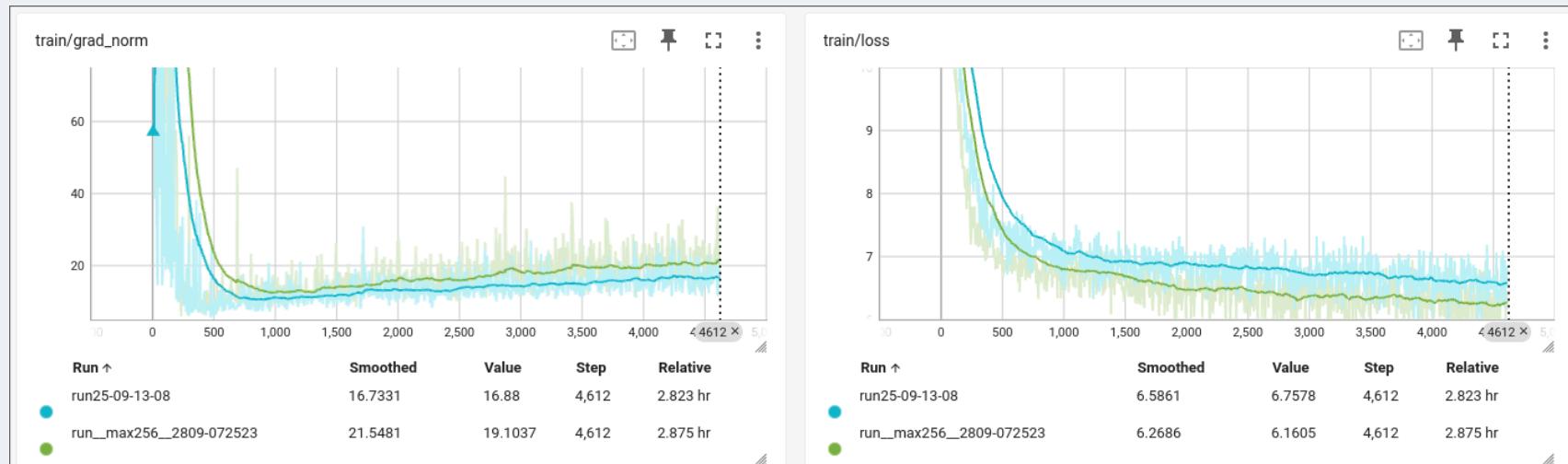


Figure 1: Green with more target modules

4. Ill conditioned Dataset ?..

mon Sep 30 11:33:01 AM CEST 2024

1. visual question answers are'nt supposed to give out long answers[even openended]?
 - Most of the model i read like TUM max token generation
 - With respect to the maximum text length accepted by CLIP, which is 76, **we trimmed any longer captions**, while zero-padding shorter ones. We refer to the resulting fine-tuned model as PubMedCLIP. PubMedCLIP was trained for 50 epochs

— Pubmedclip

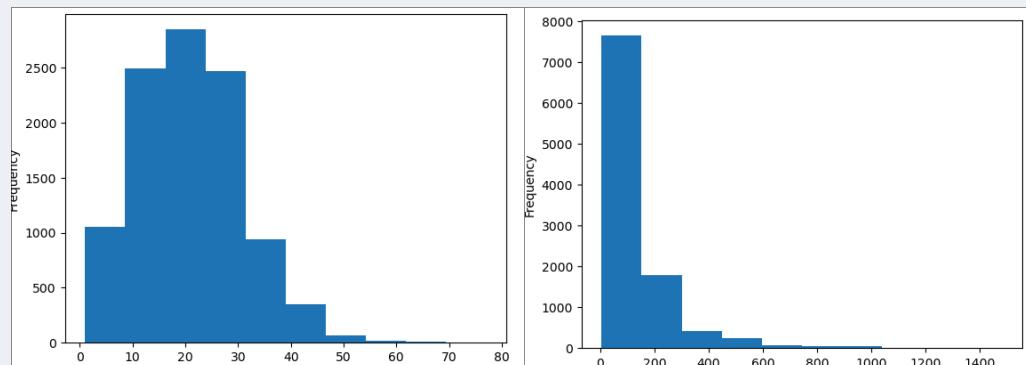


Figure 2: Left Generated by Llama [alignment] , Right Directly from captions [pre-training]

1. But good signs of getting better results than Pubmedclip
 - **we can accept a higher a context window size** : most of the results use 512. i am doing currently at 256

5. Epoch 1 results with context window 256

Monday Sep 30 11:33:01 AM CEST 2024

answer en what abnormalities does the patient have? The patient has brain swelling left atrium ventricle deformity left ventricle atrophy and the midline occipital a small ventricle and ventricle.



{"U_id": "MPX1014", "image": "MPX1014_syncpic41745", "caption": "Axial T2 FLAIR demonstrates hyperintensity in the bilateral thalamus, and periaqueductal grey matter. Post contrast imaging reveals enhancement in the periaqueductal grey matter, and bilateral mamillary bodies.", "type": "MRI", "case_info": {"Title": "Wernicke encephalopathy", "History": "52 year old woman with a history of breast cancer, on chemotherapy. Poor oral intake and now stuporous.", "Findings": "T2 hyperintensity within the both thalamai\n T2 hyperintensity of the periaqueductal grey matter\n Enhancement of both mamillary bodies\n Enhancement of the periaqueductal grey matter", "Differential Diagnosis": "bilateral thalamic glioma\n metabolic/toxic processes (Wernicke encephalopathy, Leigh disease, Wilson disease)\n infectious agents (Japanese and West Nile encephalitis, Creutzfeldt-Jakob disease)\n vascular lesions(deep venous thrombosis, top of the basilar syndrome, infarction of the artery of Percheron)", "Case Diagnosis": "Wernicke encephalopathy", "Diagnosis By": "Pathology, laboratory, and Imaging", "Treatment & Follow Up": "Patient was given thiamine, but expired", "Discussion": "Limited differential diagnosis - MR imaging is the modality of choice. Differential diagnosis can be narrowed by the imaging appearance in combination with the presence or absence of other regions of involvement, & the patient's history. The differential includes primary neoplastic processes (bilateral thalamic glioma), metabolic/toxic processes (Wernicke encephalopathy, Leigh disease, Wilson disease), infectious agents (Japanese and West Nile encephalitis, Creutzfeldt-Jakob disease), & vascular lesions(deep venous thrombosis, top of the basilar syndrome, infarction of the artery of Percheron)", "topic_info": {"Title": "Wernicke encephalopathy", "Disease Discussion": "Severe neurologic disorder resulting from dietary vitamin B1 (thiamine) deficiency - frequently associated with chronic alcohol abuse. Thiamine is an essential coenzyme in intermediate carbohydrate metabolism, and is also an osmotic gradient regulator. Deficiency may cause swelling of the intracellular space along with a local disruption of the blood-brain barrier. Wernicke encephalopathy is associated with chronic alcohol abuse, but can be caused by other disorders such as: gastropathy for obesity, psychogenic refusal of food, hyperemesis gravidarum, anorexia nervosa, voluntary food starvation, and parenteral therapy. The healthy body has a reserve of thiamine sufficient for up to 18 days, so any condition resulting in thiamine deficiency lasting 2 to 3 weeks may be an etiology\n It is a medical emergency treated by intravenous administration of thiamine. Clinical triad of alteration in consciousness, ataxia, & ocular dysfunction\n Imaging findings demonstrate symmetric T2 hyperintensity within the mamillary bodies, medial thalamus, tectal plate, & periaqueductal grey\n Post contrast enhancement may or may not be seen. Thiamine is an osmotic gradient regulator, & deficiency may result in disruption of the blood brain barrier - leading to enhancement\n Reduced diffusion on DWI may be seen. This may be related to ischemic-like changes occurring within the thalamus.", "ACR Code": "1.5", "Category": "Toxic (see also Metabolic)", "Keywords": "Wernicke encephalopathy\n Mamillary body\n Thiamine", "Reference": "American Journal of Neuroradiology 29:164-169, January 2008"}}

Figure 3: cherry picked good answer

answer en what abnormalities does the patient have? A 1 mm tumor of the left cecus and a and 1. A small renal cell nodule. the colon has a heterogeneous mass right cecum. cecus cecum., cecal mass. large left lower stomach. the colon. is abnormal and the cecum. There is also. the in the right cecum. cecum. The cecum. with a mass in bowel.,The colon.al has a right cecum.82 with the large cecum. cec. is small cecum. cec. and the large pelvis is. The cecum. cecus is to the colon. cec. is cec . cec. within the cec. cec. cec. cec. has. A is cec. in the cec. cec. is cec. cec. has. This is is colon. cec. as the. The cec. cec. cec. as cec. cec. to a large bowel. as. of a cec. cec. (cec. right cec. cec. cec. for. cec. cec. in. cec. cec. cec. cec.) cec. cec. cec. cec. cec. cec.



{'U_id': 'MPX1009', 'image': 'MPX1009_synpic46295', 'caption': 'Bladder is prominent with mildly thickened wall. There is a small posterolateral diverticulum on the rightward aspect.', 'type': 'TAC', 'case_info': {'Title': 'Bladder Diverticulum', 'History': '73-year-old male with hematuria and numerous white blood cells found on UA', 'Exam': 'N/A', 'Findings': 'Bladder with thickened wall and diverticulum on the right. Diverticulum is mostly likely secondary to chronic outflow obstruction.\n\nProstate enlargement.', 'Differential Diagnosis': 'Bladder Diverticulum', 'Case Diagnosis': 'Bladder Diverticulum', 'Diagnosis By': 'N/A'}, 'topic_info': {'Title': 'Bladder Diverticulum', 'Disease Discussion': 'Bladder diverticula most often occur as a result of outlet obstruction. Occasionally, a congenital weakness in the bladder wall adjacent to the ureteral orifice results in a diverticulum. This is termed a "Hutch" diverticulum.\n\nIn children, outlet obstruction causing a diverticulum is rare and can be seen with urethral valves. In men, diverticula are associated with outlet obstruction from urethral stricture, prostatic hypertrophy, prostatic carcinoma etc. acquired diverticula are rare in women.\n\nDiverticula usually occur on the lateral bladder walls, rarely the dome. They are often multiple. Large diverticula often displace the bladder and or ureters. Diverticula can have wide or narrow necks. The wide necked variety empty urine readily. The narrow neck type are slow to empty and therefore are more likely to have urinary stasis.\n\nInfection, tumor and stone formation can occur as a result of urine stasis within a diverticulum. Tumor formation in a diverticulum is more likely to spread beyond the bladder because the diverticulum wall consists only of urothelium without muscle.\n\nBladder diverticula can be evaluated with excretory urography, ultrasound, CT and cystoscopy.\n\nRef:\nDunnick, R., McCallum, R., Sandler, C., Textbook of Uroradiology', 'ACR Code': '8.9', 'Category': 'Diverticulum'})}

Figure 4: cherry picked bad answer

6. Scaling Law

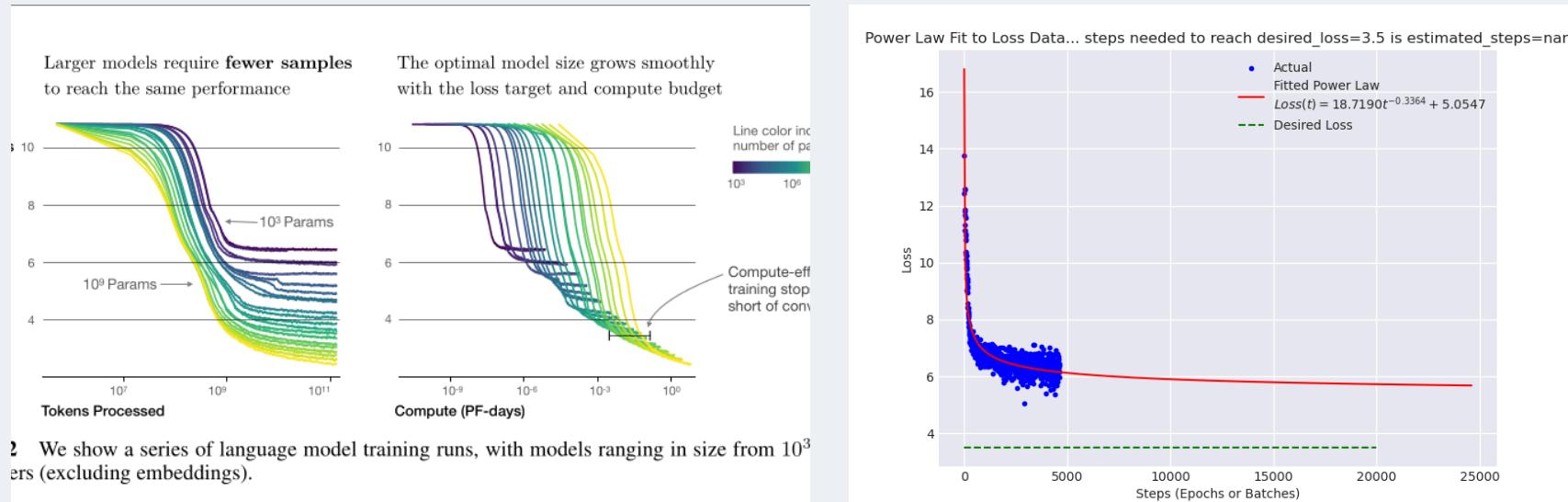


Figure 5: Scaling law

$$y = a(t^{-b}) + c$$

1. We will never get to the desired loss with our current hyperparameters/dataset
2. shows theoretical as 5.05 ...
3. we will do this again.. [possibly miscalculated]

7. Done

<|end of notes|>

NOTES

Purpose of report:

xfolds:

- give effective making techniques for longer sequence lengths (compare)
- scaling law to give appropriately sized parameter model
- to multiple adapters training or to single encoder decoder training (but this with longer sequence)
- compare results with previous similarly sized models.

MedClip https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch_32

PubMedCLIP was trained on the [Radiology Objects in COntext \(ROCO\)](#) dataset, a large-scale multimodal medical imaging dataset. The ROCO dataset includes diverse imaging modalities (such as X-Ray, MRI, ultrasound, fluoroscopy, etc.) from various human body regions (such as head, spine, chest, abdomen, etc.) captured from open-access [PubMed](#) articles.

PubMedCLIP was trained for 50 epochs with a batch size of 64 using the Adam optimizer with a learning rate of 10⁻⁵. The authors have released three different pre-trained models at this [link](#) which use ResNet-50, ResNet-50x4 and ViT32 as image encoders. This repository includes only the ViT32 variant of the PubMedCLIP model.

- Repository: [PubMedCLIP Official GitHub Repository](#)
- Paper: [Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?](#)

- Current approaches for this multimodal task adopt deep neural encoders to interpret the image and the question and then pick a corresponding answer. They typically consist of four main components: a visual encoder, question encoder, attention-based fusion of vision and text features, and an answer classifier Vu et al. [2020], Zhan et al. [2020], Nguyen et al. [2019], Pan et al. [2021], Liu et al. [2021b].

[Ideas]

They already used 2 different Encoders embedding for text and images[specialized].

so now we can just try to make 1 unified encoder and do ablation

[we aren't doing ultrasound] so would it be fair comparison. so maybe just use their tokenizer.?

other notes

- find benchmark datasets if you don't use ultrasound , flourescopy

GPU

The experiments were conducted using
one 32 GB GPU (Nvidia DGX1 8x Tesla V100) in an OKD 4.6 cluster
under the Maastricht University Data Science Research Infrastructure

To compute the performance in petaflop days (PF-days) for an Nvidia DGX-1 with 8x Tesla V100 GPUs, we can break it down as follows:

1. Understand PFLOPS for Nvidia DGX-1 with 8x V100:

A Tesla V100 (32 GB) GPU can deliver up to 15.7 teraflops (TFLOPS) of single-precision (FP32) performance.

The DGX-1 has 8 V100 GPUs, so the total peak single-precision performance of the DGX-1 is:

Total TFLOPS=15.7×8=125.6 TFLOPS

Total TFLOPS=15.7×8=125.6 TFLOPS

Convert TFLOPS to petaflops (PFLOPS):

Total PFLOPS=125.61000=0.1256 PFLOPS

Total PFLOPS=1000125.6=0.1256 PFLOPS

2. Time Calculation for PF-days:

1 PF-day is equal to 1 petaflop sustained for 1 day (i.e., 24 hours or 86,400 seconds).

Now, you can calculate how many PF-days the DGX-1 can achieve in a given time period. For 1 day, the DGX-1 can achieve:

PF-days=0.1256 PFLOPS×1 day=0.1256 PF-days/day

This means the Nvidia DGX-1 with 8x Tesla V100 GPUs can deliver 0.1256 petaflop-days of computation in one day.

from 4m

- All modalities are mapped to sets or sequences of discrete tokens (indices of a vocabulary) through the use of modality specific tokenizer
- captions and bounding boxes are both treated as text and tokenized by wordpiece

- Encoder is a std transformer but features modality-specific learnable input embedding layers to map token indices to vectors .
 - Non learnable sin-cos position embedding.
-
- The decoder handles tokens from both dense image-like and sequence-like modalities, with each type requiring a different approach

Quick Summer Summary 😊

I read an article from Korean Journal of Radiology which had some nice things to say about multimodal models on radiology and health care.

The issues it mostly talks about were :

- achieving generalizability of AI models
- establishing the explainability of the decision-making process

-> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613849/>

To overcome these issues, some of the various possible solutions are as follows:

inclusion of training with longitudinal and multimodal datasets, dense training with multitask and multimodal learning, new generative models including anomaly detection, XAI

--- From Summary and conclusion

I Mostly read on Collecting data and tokenization , since that will be our first steps.

For Data:

- Medpix2.0
 - 12,000 patient case scenarios, 9,000 topics, and nearly 59,000 images.
 - But Far less than what 4m pretrains on which is 12M (CC12M) .
 - But Medpix claims : CLIP succeeds in achieving competitive performance in zero-shot contexts on a wide range of classification datasets by learning the relationships between images and their textual descriptions entirely trained on their own dataset.

For tokenization

- Here 4m does different than almost every other paper
- 4m has 1 unified embedding for all modalities
- But Llama , Medpix and others have different image and text encoder. [so as many embeddings and cross attention as modalities]
- Having different embedding would allow us to use specialized tokenizer like BioBert.[<https://doi.org/10.1093/bioinformatics/btz682>, arXiv:1901.08746 [cs]]
- But 1 unified embedding would be great to plot after pretraining
- So I am not really sure of what tokenization technique i should use for now.

Extra Ideas

Just so that its not complete replication of any 1 paper. I was thinking of making some changes by taking pieces of others architectures so that i could do ablation studies [Thats a research gap i think].

- like 4m does not use grouped query attention (which is supposedly faster with same resolution as Multihead attention) (unlike Llama 3.1).
- 4m does a lot of masking (Massively Masked Multiple Modality) but does not talk about efficiency of this masking. A lot of other paper CLIP [<https://arxiv.org/pdf/2103.00020.pdf>] and SIGLIP [<https://arxiv.org/abs/2303.15343>] which newer models are trained on ; go in a lot of detail.
 - I did some triton and cuda programming over the summer . Maybe i can try writing their code in cuda/triton to show if their methods are as efficient and fast as others or not. (But this seems slightly out of scope for the internship plan we submitted!)