Castrellón Carrillo Francisco Alfredo
Robles Jiménez Luis Eduardo

Medical Image Captioning

INTRODUCCIÓN

I. Introduction

Image captioning is the task of automatically generating an accurate text description based on an image. Among its multiple uses, generating medical diagnosis using images has become one of the most important nowadays. Medical images are essential tools for clinical diagnosis since they allow doctors to diagnose various medical conditions; however, because of the complexity of these images, it requires experienced professionals to deliver a correct diagnosis and it's a very time consuming task. X-ray imaging is one of the most used methods because of its affordable and accessible nature and it is used to diagnose not only broken bones but a wide variety of lung diseases, making it hard to generalize exactly what to look for at first, since different diseases may have similar effects with very subtle differentiators, and this problem is common among medical imagery. Thus, there is a current need for a tool that reliably guides doctors in the areas of images they should focus on, as well as in interpreting and summarizing possible disease symptoms based on images; so, by fighting the world's growing shortage of medical doctors, the process could become more efficient and effective, more people could be diagnosed, and ultimately more illnesses could be treated. Because of the importance of this problem, there has been a lot of studies and approaches to develop tools that help professionals with this task. From simple image segmentation to detect anomalous sections of an image to deep learning models to predict illnesses based on images; the specific application covered in this paper is automatic caption generation for medical images, which is a core technology for

automated diagnosis where the goal is to generate accurate, relevant and coherent descriptions of the images using medical terminology to provide doctors valuable information in a quickly manner. In general, these systems work by extracting features from images with a specific method so they can be later used to generate captions with a text generating technique. This is a big challenge because of the nature of this area, since we are diagnosing people, the terminology needs to be very precise and the description coherent so it does not confuse the user or cause an incorrect diagnosis. Nowadays due to the success of deep learning in image detection and text generation it is being researched as a solution to this problem.

This paper is intended to be a participating approach in the image captioning task of the 2023 ImageCLEF campaign. And the contributions are:

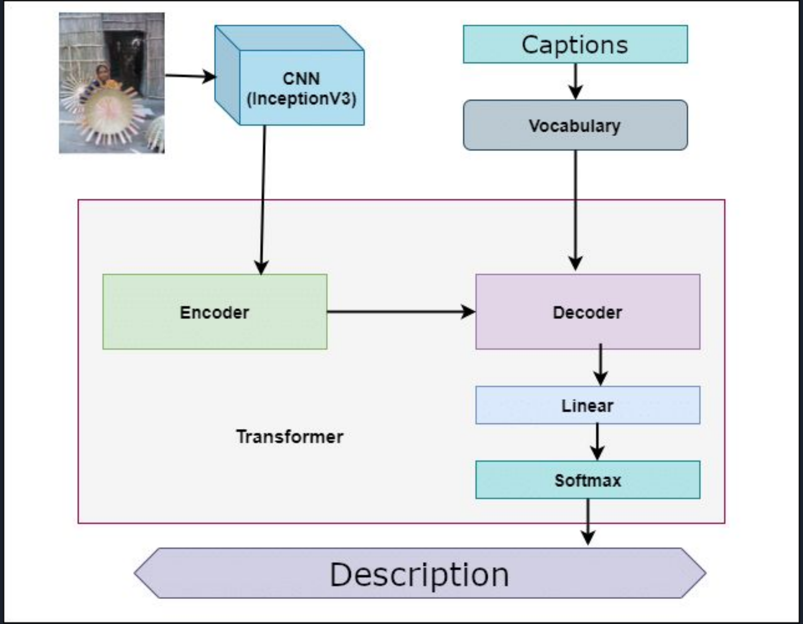
- Exploration of two different feature extraction approaches to compare their results against their demand of training resources.
- Train an existing model with a different dataset to see how it behaves with medical imagery.

ESTADO DEL ARTE

II. Related Work

Recently, multiple encoder-decoder systems have been implemented. Usually, convolutional neural networks (CNN) perform encoder tasks to extract visual information as features, and recurrent neural networks (RNN) are used as decoders to generate the final text. Other works use RNNs to transform extracted features into shape-related information and then use long short term memory (LSTM) to get the final description [1]. Currently this is the basis for medical image captioning, but there are many works that by introducing different modifications achieve better results in specific areas, for example, Yuxuan Xiong et al. [5] in "Reinforced Transformer for Medical image Captioning" proposes a hierarchical transformer based model consisting on an image encoder that extracts heuristic visual features by a bottom-up attention mechanism and a non-recurrent captioning decoder that improves computational efficiency by parallel computing and achieving a BLEU-1 performance increase of 50% compared contemporary methods. Other state-of-the-art methods use attention and transformer models [2, 3], like "Medical image captioning via generative pretrained transformers" by

Alexander Selivanov et al. [3] where they use the Show-Attend-Tell model and the GPT-3 model to generate a textual summary of radiology records containing essential information about the pathologies found, the location and 2Dheatmaps that localize each pathology and with that combination of language models they outperformed other models, introducing a new preprocessing pipeline that allows the obtention of higher metrics.



METODOLOGIA

Vocabulary cleaning

```python
def remove_punctuation(text_original):
    text_no_punctuation = text_original.translate(string.punctuation)
    return(text_no_punctuation)

def remove_single_character(text):
    text_len_more_than1 = ""
    for word in text.split():
        if len(word) > 1:
            text_len_more_than1 += " " + word
    return(text_len_more_than1)

def remove_numeric(text):
    text_no_numeric = ""
    for word in text.split():
        isalpha = word.isalpha()
        if isalpha:
            text_no_numeric += " " + word
    return(text_no_numeric)

def text_clean(text_original):
    text = remove_punctuation(text_original)
    text = remove_single_character(text)
    text = remove_numeric(text)
    return(text)

for i, caption in enumerate(train_df.caption.values):
    newcaption = text_clean(caption)
    train_df["caption"].iloc[i] = newcaption
```

```
Clean Vocabulary Size: 26595
```

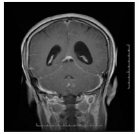
```python
clean_vocabulary = []
for txt in train_df.caption.values:
    clean_vocabulary.extend(txt.split())
clean_vocabulary = set(clean_vocabulary)
print('Clean Vocabulary Size: %d' % len(set(clean_vocabulary)))
print(clean_vocabulary)
```

✓ 0.1s

```
Clean Vocabulary Size: 26595
{'Troendle', 'decitabine', 'interatial', 'Cx', 'PCIs', 'wallstent', 'Abdominopelvic',
```

Dataset

Postop 22-month CT scan (sagittal): Posteriorly the graft seated in a sound bone

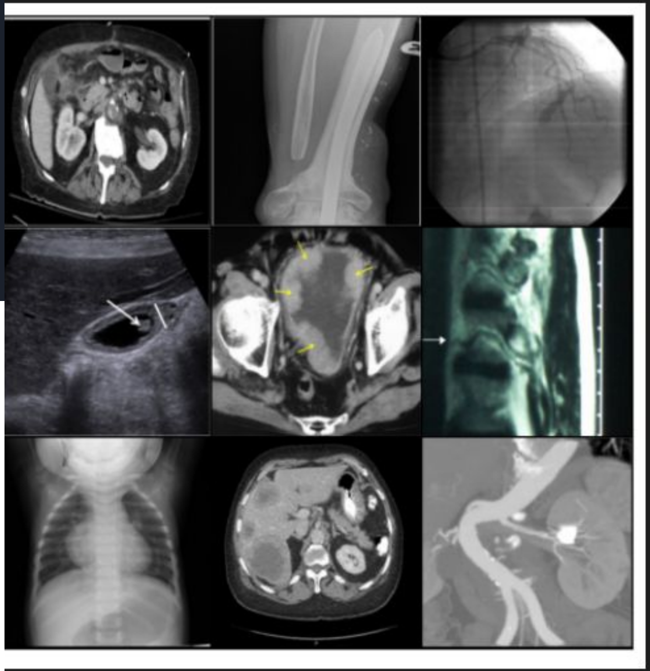


Enhanced magnetic resonance imaging of head revealed bilateral cerebral and cerebellar hemispheres abnormal meningeal enhancement.

Enhanced magnetic resonance imaging of spinal cord delineated multiple enhancement nodules in spinal cord, cauda equina, and cristae membrane (arrow).

Sagittal T2-SPAIR image illustrating the "fluid sign (arrow)" in the acute osteoporotic compression fracture.

CT demonstrating partially obstructed airway.CT: computed tomography.



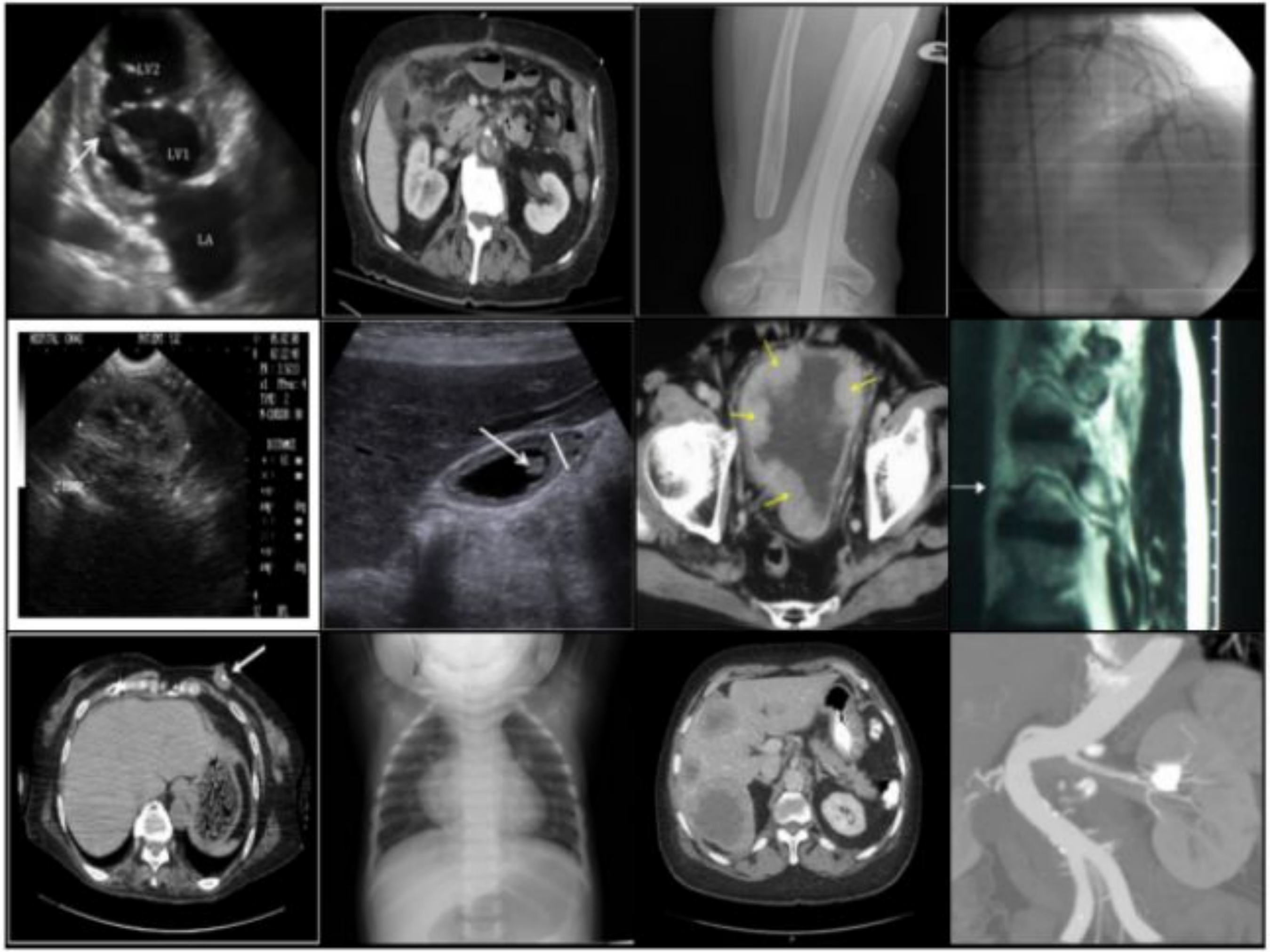
Vocabulary Size: 66897

FOLLOW UP
Data Loader

```python
class ImageCaptionDataset(Dataset):
    def __init__(self, csv_file, root_dir, transform=None):
        self.annotations = pd.read_csv(csv_file, sep = "\t")
        self.root_dir = root_dir
        self.transform = transform
    def __len__(self):
        return len(self.annotations)
    def __getitem__(self, index):
        img_id = self.annotations.iloc[index, 0]
        img_name = img_id + ".jpg"
        img_path = os.path.join(self.root_dir, img_name)
        #image = Image.open(img_path).convert("RGB")
        image = plt.imread(img_path)
        if self.transform is not None:
            image = self.transform(image)
        caption = self.annotations.iloc[index, 1]
        return image, caption
```

```python
# Create train, validation and test datasets
train_dataset = ImageCaptionDataset(csv_file='data/dataset/train_labels.csv',
root_dir='data/dataset/train_images/train', transform=transform)
val_dataset = ImageCaptionDataset(csv_file='data/dataset/valid_labels.csv',
root_dir='data/dataset/valid_images/valid', transform=transform)

# Create data loaders for each dataset
train_loader = DataLoader(train_dataset, batch_size=16, shuffle=True)
val_loader = DataLoader(val_dataset, batch_size=16, shuffle=True)
```


LV2
LV1
LA

Feature extraction

Inception 3 (CNN)

```python
def load_image(image_path):
    img = Image.open(image_path)
    transform = transforms.Compose([
        transforms.Resize(299),
        transforms.CenterCrop(299),
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406],
                             std=[0.229, 0.224, 0.225])])
    img = transform(img).unsqueeze(0)
    return img, image_path
# Load the inception v3 model
image_model = models.inception_v3(pretrained=True)
# We're not training so the gradients are turned off for the model
for param in image_model.parameters():
    param.requires_grad_(False)
# Move the model to the GPU
image_model = image_model.to(device)
image_model.fc = torch.nn.Identity()
image_features_extract_model = image_model.eval()
```

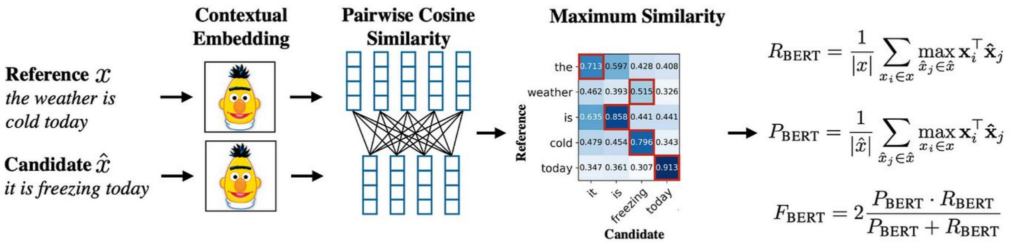
```python
for img, path in tqdm(image_dataloader):
    with torch.no_grad():
        img = img.to(device)
        batch_features = image_features_extract_model(img)
        batch_features = batch_features.cpu().detach()
        for bf, p in zip(batch_features, path):
            path_of_feature = p.split("/")[-1]
            path_of_feature = path_of_feature.split(".")[0]
            np.save("features/" + path_of_feature, bf.numpy())
```

✓ 6m 0.3s

```
100%|██████████| 625/625 [06:00<00:00,  1.73it/s]
```

Metrics

Introducing **BERTScore**



Source: Bertscore: Evaluating text generation with bert

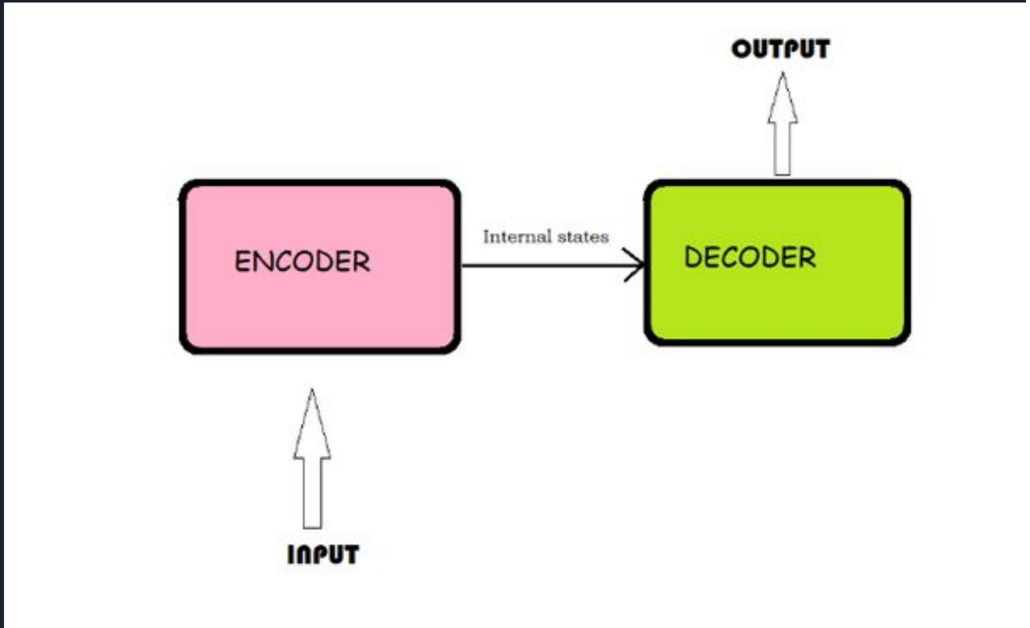
Code for Bertscore is available at https://github.com/Tiiiger/bert score

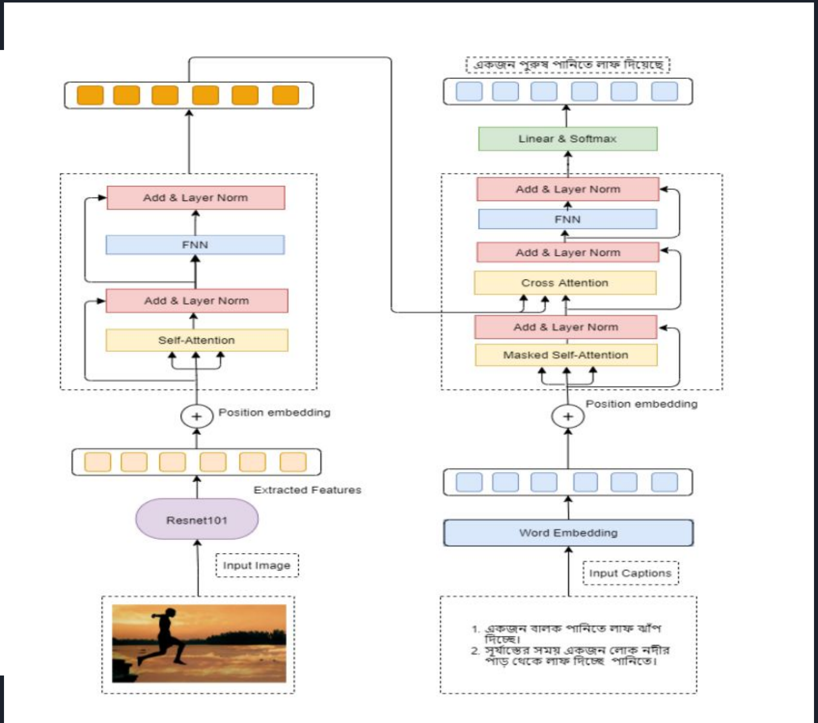
BERTScore

- Contextual embedding
- More robust
- Longer memory
- Rescaling

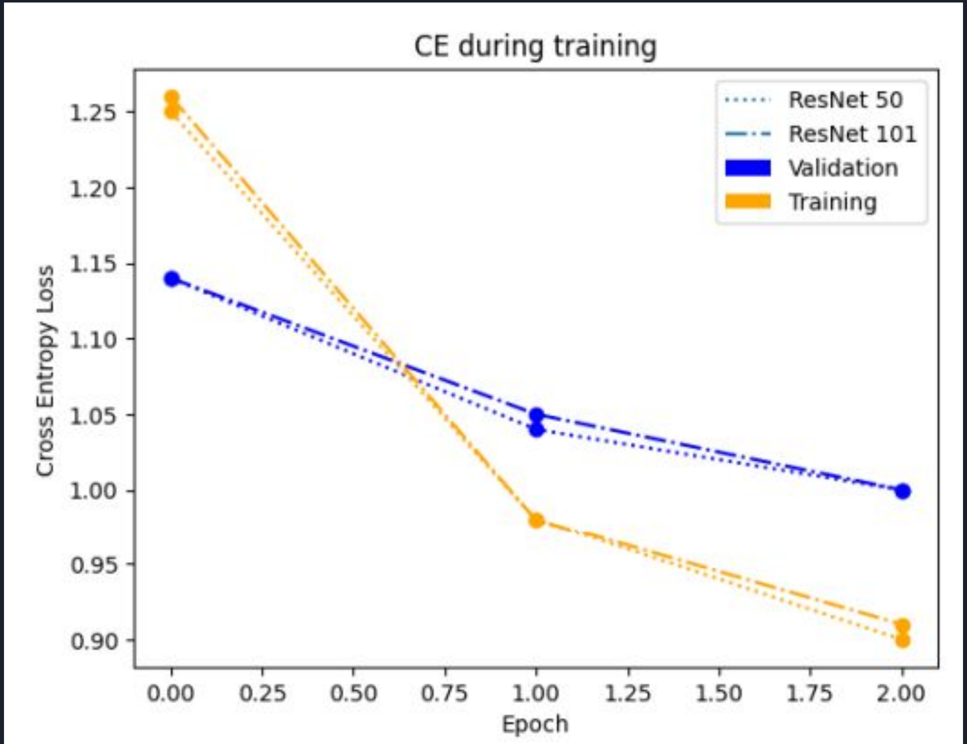
Model type: `microsoft/deberta-xlarge-mnli`

Encoder-Decoder





Experiments-Performance



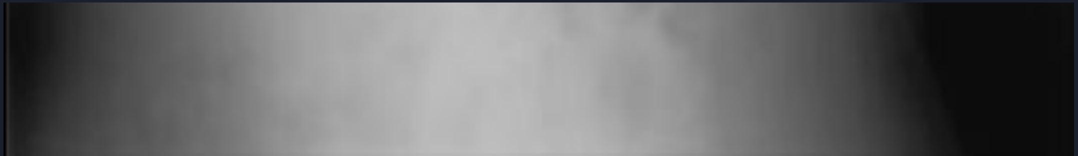
Experiments-Metrics

Encoder Model	Minutes per epoch	# of parameters	BERT F1
ResNet-50 @ 3	197	120M	0.5949
ResNet-101 @ 3	244	138M	0.6111

TABLE I

FOLLOW UP
Metrics

Postoperative chest radiograph showing normal left clavicle



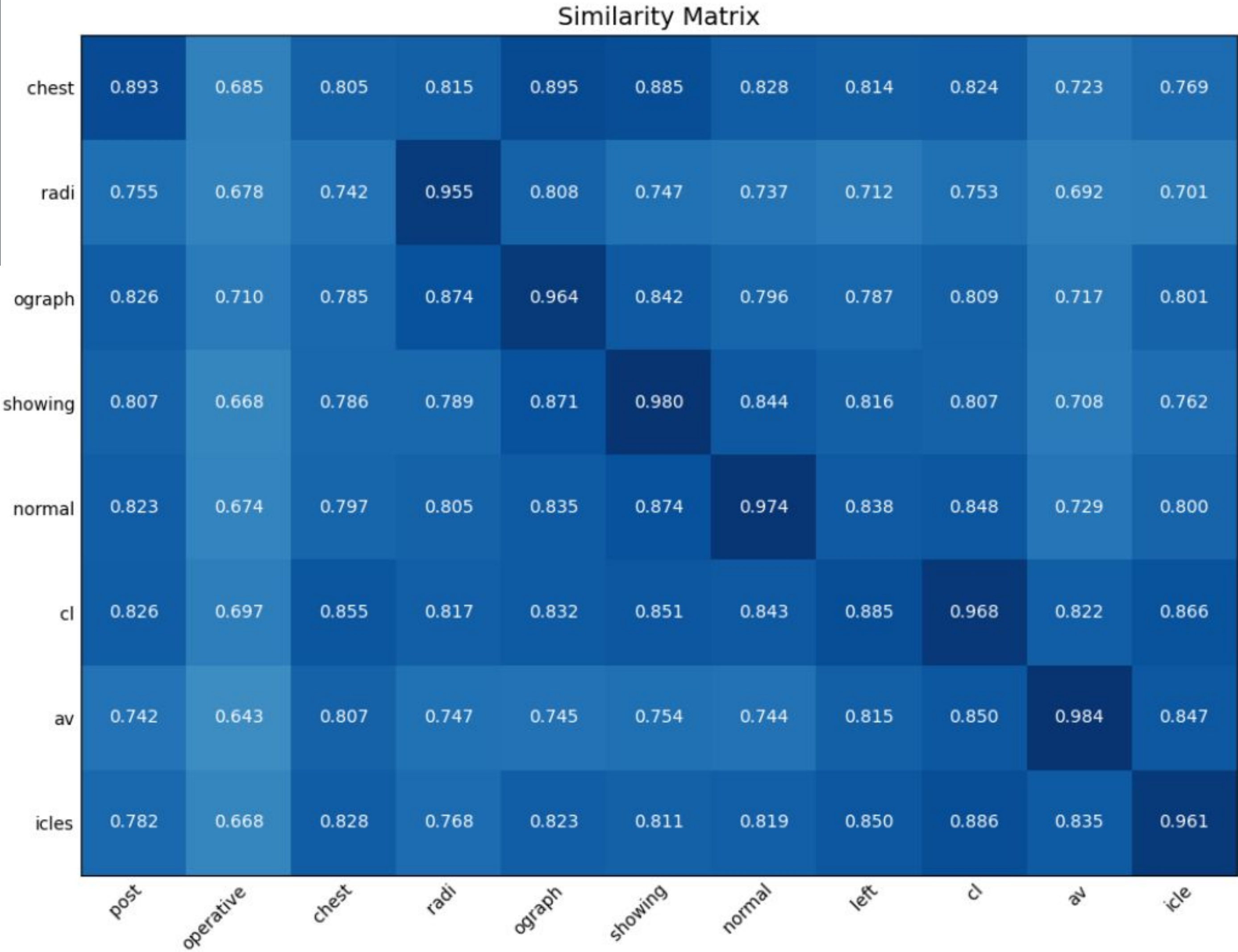
FOLLOW UP Metrics

Reference

postoperative chest radiograph showing normal left clavicle

Candidate

chest radiograph showing normal clavicles



FOLLOW UP Metrics

Reference

postoperative chest radiograph showing normal left clavicle

Candidate

chest radiograph showing normal clavicles

FUTURE TESTS

We are testing Multihead Attention model, we will try different models

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concate}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^{O},$$

$$\mathrm{wherehead}_i = \mathrm{Attention}\left(QW_i^{Q}, KW_i^{K}, VW_i^{V}\right).$$

Different feature extraction models

Fine tuning Generative Pre-trained Transformer (GPT) models

Conclusión

Image	Reference	Predicted
	Posttreatment CT showing no residual pancreatic tumor	ct scan of the abdomen showing a large mass in the right kidney.
	An AP radiograph showing the healed osteotomy	x - ray of the left knee showing a well - defined radiolucent lesion in the left femur.

TABLE III

- Hardware consumption
- backbone not the main problem