NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

MSc: DATA SCIENCE AND INFORMATION TECHNOLOGIES

Deep Neural Networks Assignment

Biomedical Image Captioning

Chris Morfopoulos (7115152100011)

Supervisors: Haris Papageorgiou

Athanasia Kolovou

Table of Contents

1.Abstract

In the last years, the deep learning task of Image captioning was very challenging. Image captioning is a combination of Computer Vision and Natural Language Processing, which are the two 2 sub main areas of deep learning. Essentially, Image captioning is an end-to end sequence to sequence embedding task where, image pixels are input sequences and captions are describing the images as a desired output. In this way, with respect to an image, a model can generate an output as a label or caption. The process of image captioning is very useful in many sections nowadays, especially in the medical sections where a model can briefly describe in few words the condition of a patient. However, the implementations of Image captioning, with older architectures such as Convolutional Neural Networks and LSTMs or RNN, had a very low performance mainly due to the limitations of the LSTM architecture. The attention mechanism along with the Transformer architecture which was introduced in "**Attention Is All You Need**" paper **[1]** had great impact both in NLP with Language Models such as Bert and in Computer Vision with image recognition models such as ViT. In this way, applying these new architectures in the image captioning task can improve the overall performance of the models. In this paper, we will explore not only the old but also the new architectures of Image captioning and monitor the performance of the respective models.

2.Introduction

As we mentioned, Image captioning is based on images with a main task to generate a proper caption with respect to the image. This can be easily adapted for medical X-ray images to predict a summary describing the status of a patient. This can be incorporated in many medical institutions helping the medical staff with their daily workload. The main concept in image captioning architecture, is separated in two main steps. In the first step the goal is to extract the main features of an image and in the second step, with these features, to generate an output that generally describes properly the image.

The conventional architecture of Image captioning was based on Convolutional Neural Networks and LSTMs/ RNNs. The features of an image were extracted by pretrained CNNs and the LSTM network generate an output with respect of the features. There were a lot of restrictions in the above architecture mainly in the LSTM or in the RNN network and this had as a result a low performance. LSTMs as sequence-to-sequence models have not only a tendency to overfit but they require a lot of time to train mainly due to the fact that have many parameters. On the other hand, the computation of RNNs is very slow and the training process can be very difficult.

A major turning point in the deep learning area was when the "**Attention is All you need**" paper was published **[1]**. In this specific paper, the authors introduced the Transformer architecture, which consists of Encoder and Decoder layers along with the attention mechanism. This architecture was originally designed for Machine Translation tasks; therefore, the inputs are pre-processed words. The attention procedure is the main core of Transformers, and its goal is to measure how much focus the model should put to certain words within a sentence. Several language models that were based on Transformer architecture came up and are considered as state of the art for NLP tasks. Some popular language models are BERT, GPT2, MT5 and Roberta.

In this way, the attention mechanism gained so popularity that the authors of **[1]** tried to apply this architecture to other tasks such as Image recognition. Obviously, the inputs of the respective architecture will be images and not words as the original Transformer had as input. In this way, the authors tried to modify the original and the already successful architecture of Transformers as less as possible. Therefore, in the "**An image is worth 16x16 words: Transformers for Image Recognition at Scale**" paper **[2]**, the Vision Transformer was introduced to the community. The Vision Transformer architecture was designed for Image classification, and it was relied on the original Transformer architecture but with some small modifications. Finally, the ViT models marked an extraordinary benchmark in classification tasks and outperformed the CNN models.

Inspired by the main structure of Transformer architecture, we can apply the **encoder – decoder philosophy** for **image captioning**. Moreover, the encoder section, can be replaced by a pretrained ViT model, which can easily extract the most important features of an image and the decoder section, can be replaced by a pretrained language model, which can also be finetuned to a specific corpus. To sum up, in this paper we implemented several experiments not only with the encoder-decoder concept but also with the conventional way of CNN and LSTM.

3.Dataset

The main training dataset that was utilized for our image captioning task is called ROCO dataset [3] and it stands for Radiology Objects in Context. The dataset contains over 81 thousand radiology images with several medical imaging modalities including Computer Tomography, X-ray, Mammography, Magnetic Resonance, Angiography and many more. **Due to the nature of medical captioning along with the medical terminology of the respective captions makes it hard to interpret the generated result**.

Fig. 3.1 A sample of Images along with their respective captions from the ROCO dataset

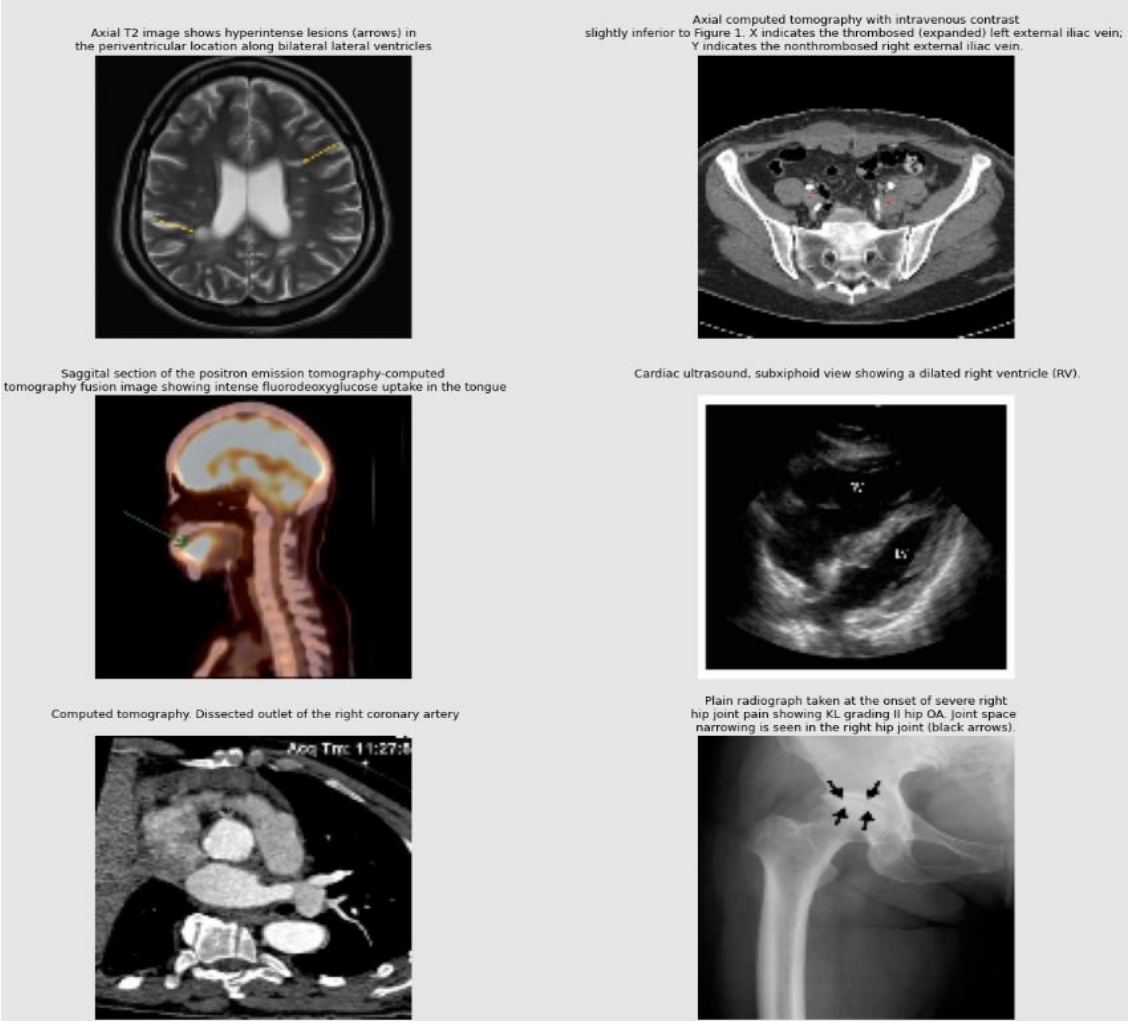


Fig. 3.2 A sample of Images along with their respective captions from the ROCO dataset

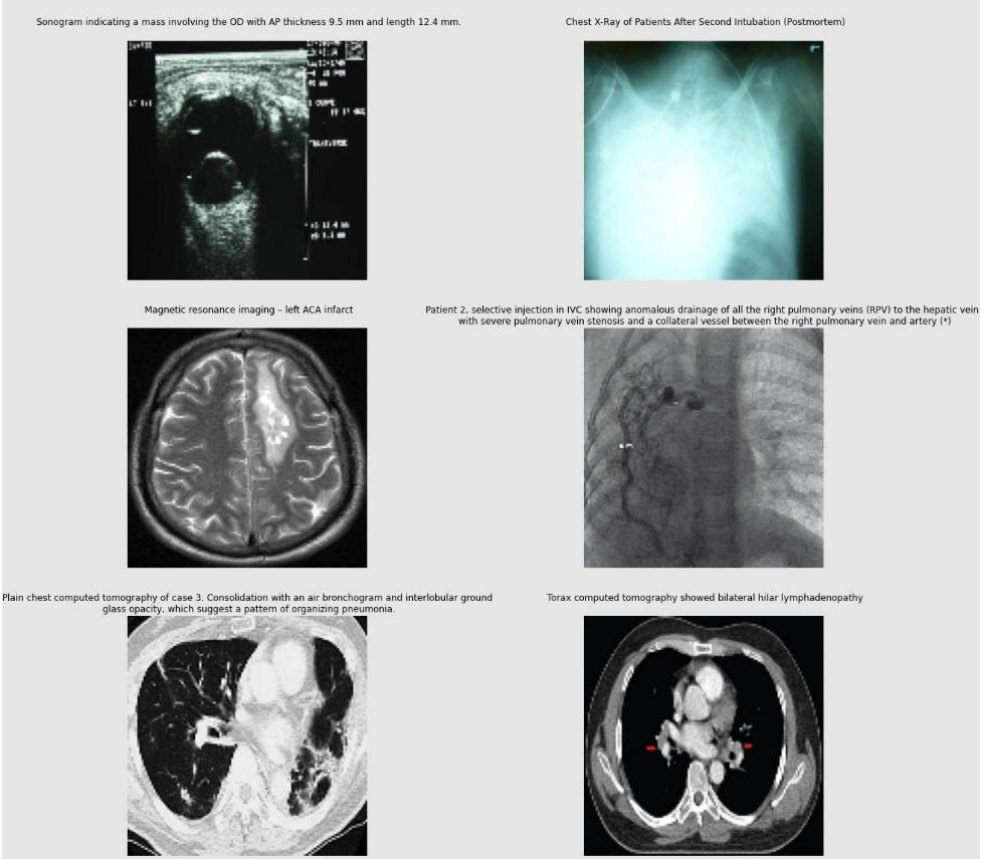


Fig. 3.3 Sample of captions from the ROCO dataset to visualize their terminology- structure

captions
FA on the left eye is normal
Impaction of the fracture during weight bearing resulted in screw joint penetration three months postoperatively.
Doppler ultrasound performed upon patient presentation demonstrating heterogeneous echotexture in both the testicle and the epididymis signifying ischemia and inflammation in the testicle and necrosis in the epididymis (red arrows). Only peripheral blood flow to the testicle is present while blood flow to the epididymis is maintained. Additionally, a significant hematoma is visualized on the anterior aspect of the testicle (green arrow).
Plain radiograph showing a lytic lesion in the right iliac wing with minimal periosteal reaction.
Sagittal transperineal sonography showing the recurrent mass in the same patient occupying the rectovaginal septum. The vagina has been filled with acoustic contrast.
Cerebral angiogram after CAS demonstrates ophthalmic artery occlusion (arrow).
CT Scan in Sagittal Section showing a coronal plane fracture of distal femur (Hoffa Fracture).
Third degree of adiposis. Clearly deteriorated dorsal transsonicity of the pancreas. Invisible splenic vein and anatomical structures located deeper. F – supraperitoneal fat
Subcutaneous epidermoid cyst at the phalanges (arrowheads). Longitudinal ultrasound showing a well-defined subcutaneous lesion with variable echogenicity (anechoic components and some internal hyperechoic debris)
PMMR of hypoxic brain changes. Axial T2-weighted PMMR image through a fetal post mortem brain, showing an example of typical low signal change in the basal ganglia which may be associated with hypoxia. Conventional PMMR cannot currently distinguish antemortem from postmortem hypoxic change
Susceptibility-Weighted Imaging at 7 T (8-channel head coil); this image represents a minimum intensity projection over a 6 mm slab. Note that both, veins and iron containing structures like the basal ganglia appear hypo intense. This image was acquired using an echo time of 15 ms and a resolution of 0.3 × 0.3 × 1.2 mm.
The primary bullet path by a full metal jacket two-shot technique
Axial CT scan of the incarcerated stomach; the first part of the duodenum is seen leaving the inguinal hernia.
Anteroposterior radiographs of a female patient who underwent bilateral Ludloff open reduction aged 5 months; a) at 1.6 years post-operatively, showing advanced subluxation of the femoral head and residual acetabular dysplasia; b) two months after bilateral Salter and femoral derotation varus osteotomies;c) after recurrence of coxa valga without aseptic necrosis at nine years of age and d) Severin group IIa of both hips at 28 years of age.
Immediate postoperative orthopentomogram
Sagittal TE weighted MR image reveal what was thought to be complete ACL rupture (arrow) was not appreciated as a complete rupture at arthroscopy. According to the arthroscopist it was a partial tear that involved approximately 75% of the ligamentous body.
These radiographs of a 39 year old male (case report 1) were made one year after open reduction and internal fixation. A good bony consolidation but an incomplete reconstruction of the joint line is visible.
PET/CT imaging. PET/CT showing that the masses displayed increased FDG uptake
Cholangiographic finding. Endoscopic retrograde cholangiography using carbon dioxide insufflation shows kinking of the common bile duct 1 cm above the proximal end of the metal stent (arrows).
CT scan 6 months after mitotane treatment showing reduction in size of the adrenal mass.
Abdominal ultrasound image in a 48-year-old female with benign lesser curvature gastric ulcer showing thickening of the gastric wall and a niche-like echogenicity (arrow), probably representing the ulcer carter.
Axial cut of CT-chest demonstrates a well-circumscribed soft tissue density (*) in the left breast measuring 3.1×3.7 cm.Abbreviation: CT, computed tomography.
Supine ventral-dorsal abdominal x-ray immediately after surgery. In the right abdominal wall is a Small Hybrid Rebound HRD. In the left side is a Dog Bone Rebound HRD.
Passing the lesion.Abbreviations: CRAN, cranial; RAO, right anterior oblique.
Increased activity in the upper left quadrant of case 3 at the 2nd hour. The activity has moved towards the inferior quadrant at the 4th hour.

However, some pre-processing techniques and assumptions were implemented in order to remove some noise from the captions. As a first step, we removed all the punctuations and converted all the captions to lowercase. Moreover, we remove all the redundant values of the captions that were lied inside of a parenthesis. We should mention that due to the restrictions of image captioning we cannot remove all the stop words from the captions, because there are essential for the Language model (or Decoder) to generate an output. Therefore, with all the

pre mentioned pre-processing actions, our caption-corpus can be utilized for training. The training dataset consist of roughly 65 thousand images and the validation as well the testing dataset consist of 8 thousand images respectively. Lastly, we created 2 main visualizations for our dataset, a world cloud with all the stop words removed and a unigram with 3-gram of the 20 most frequent word/terms.

Fig. 3.4 Word Cloud based on the captions from the ROCO dataset.

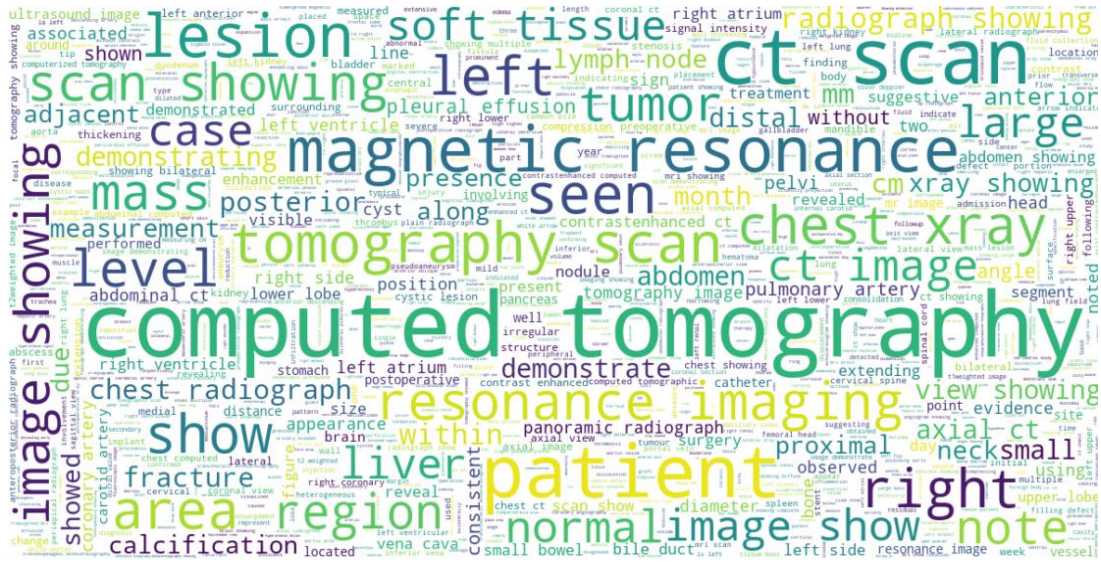
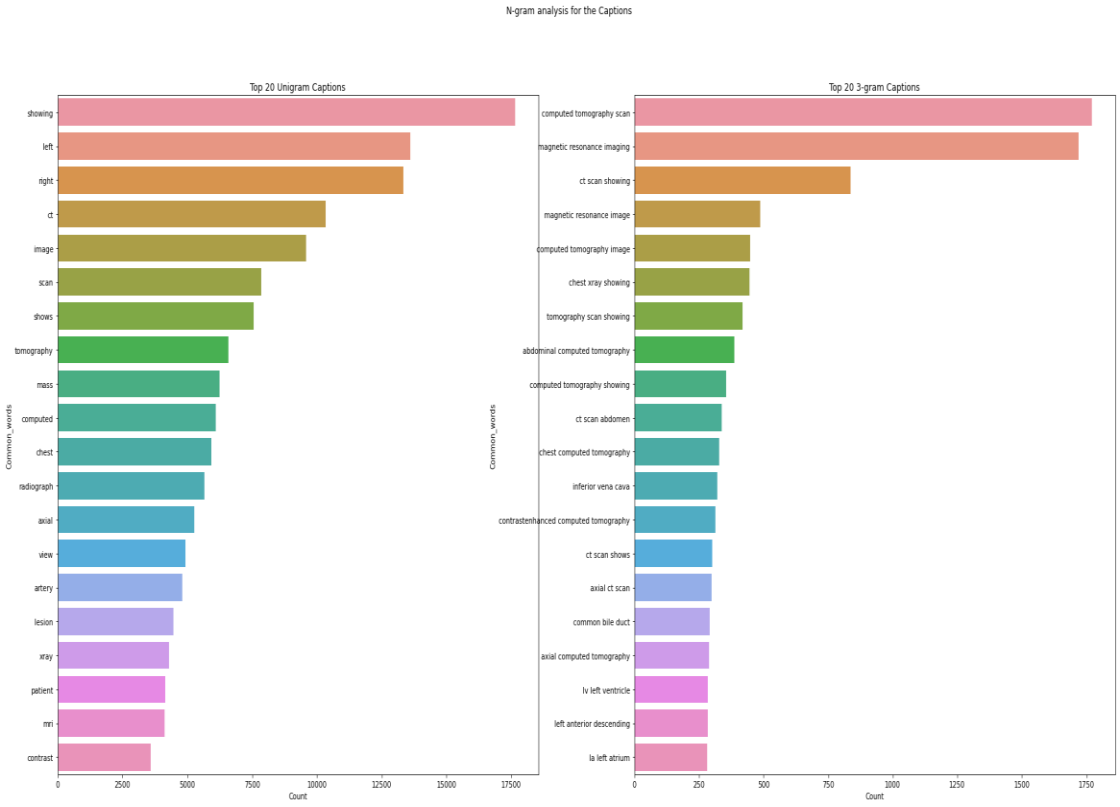


Fig. 3.5 A visualization of the top-20 most frequent words/terms in a unigram/3-gram

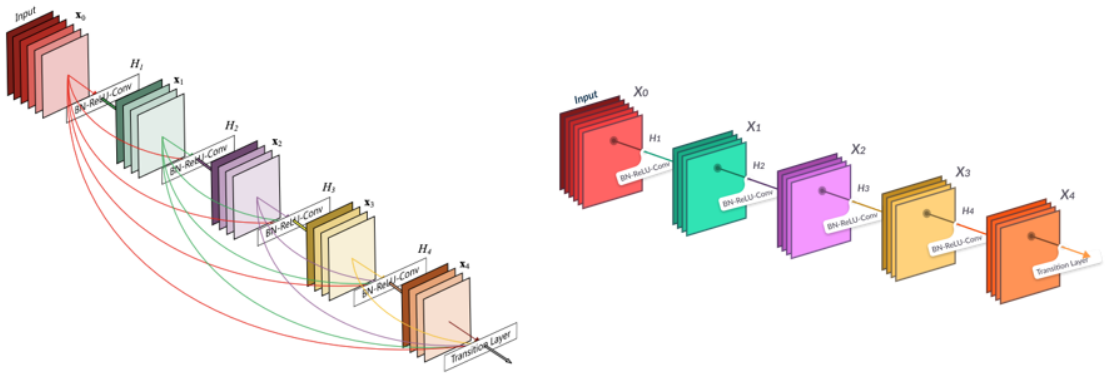


4.Architectures

As we elaborated above, the 2 main approaches for Image captioning are the conventional way with Convolutional Neural Networks and LSTM/RNN, and the contemporary way of utilizing the state-of-the-art attention mechanism with the ViT model and the transformer-based Language Models.

In our first implementation, we performed Image Captioning with the pretrained DenseNet201 as a pretrained Convolutional Neural Network and a LSTM network. The DenseNet201 model as the name suggests has 201 deep layers and is densely connected convolutional network, which is very similar to ResNet but with some fundamental differences. The ResNet is using an additive method that means they take a previous output as an input for a future layer, whereas the DenseNet takes all the previous output as an input for a future layer as shown in the below image.

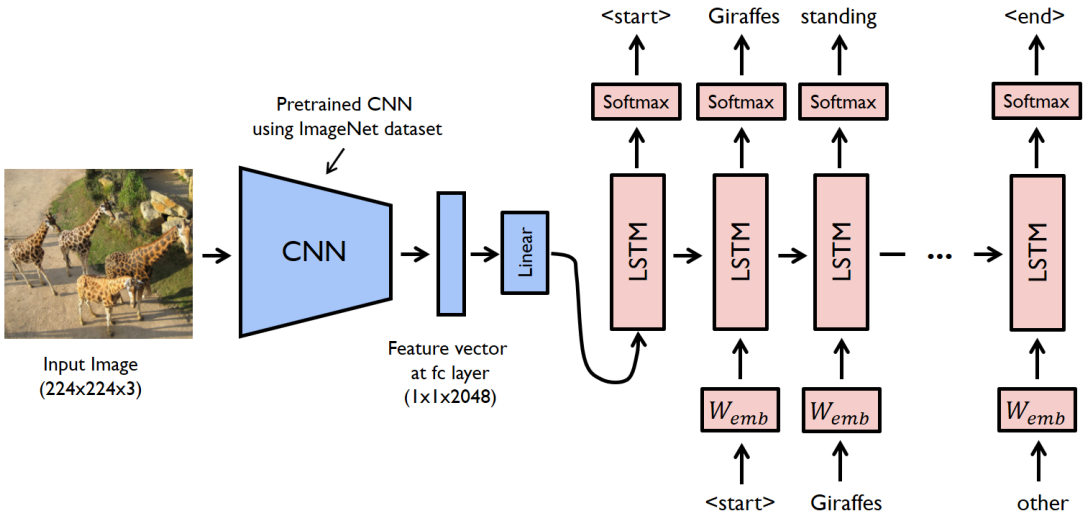
Fig. 4.1 The architecture of Dense Net in contrast to Res Net



As we mentioned above, with respect to our CNN model we extract the features from the image with a certain vector size, also known as vector embeddings. The size of our image embeddings in our case is 1920, since the Global Average Pooling layer is selected as the final layer of our model.

A very important step is to tokenize and encode the words in our captions with a Tokenizer with a certain max length and vocabulary size so we can have with these encodings a final word embedding. Additionally, we should include a token that indicated the start and the end of the sequence**. As a final step we concatenate the word embeddings with the image embeddings in order to pass them to LSTM to generate the next word, as the LSTMs are used for the text generation process**. The overall process of the Dense Net and the LSTM is shown in the figure below.

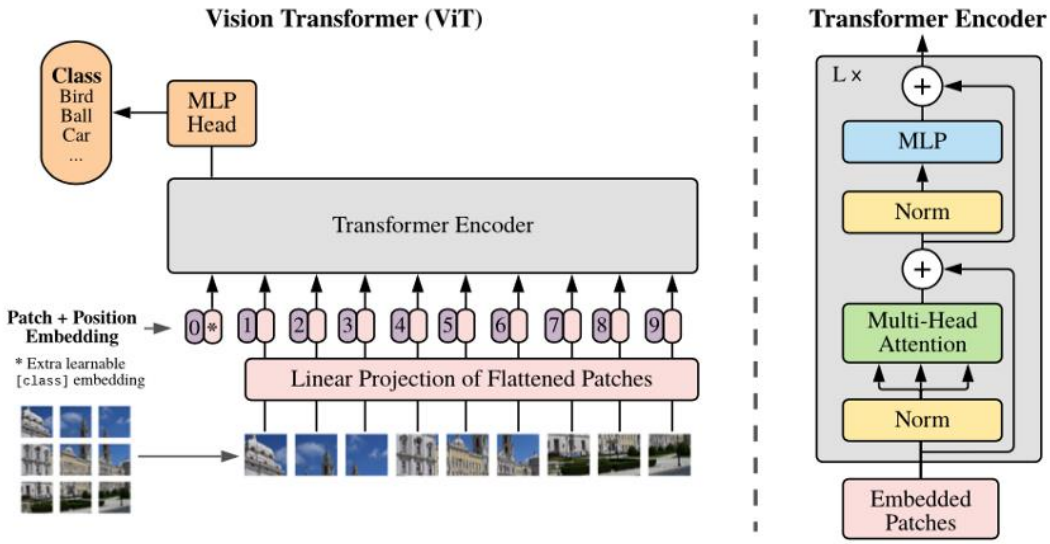
Fig. 4.2 A high level architecture of a CNN+ LSTM model



For the rest of our implementations, we implemented the Encoder-Decoder architecture as we discussed above. This architecture is strictly based in the attention mechanism. In the encoder part, where we must extract the features from the image, we applied the Vision Transformer also known as ViT.

The ViT model belongs to the Transformer Model family, and its main characteristic is that it utilizes the attention mechanism instead of convolutions to capture the features of the image. Its initial steps are to split the image into tokens **[2]** and apply therefore not only local but also global attention to the image. Finally, **the output of the N-stacked Encoders in the ViT architecture is a continuous vector representation of the inputs along with attention information**.

Fig. 4.3 The original Vision Transformer architecture



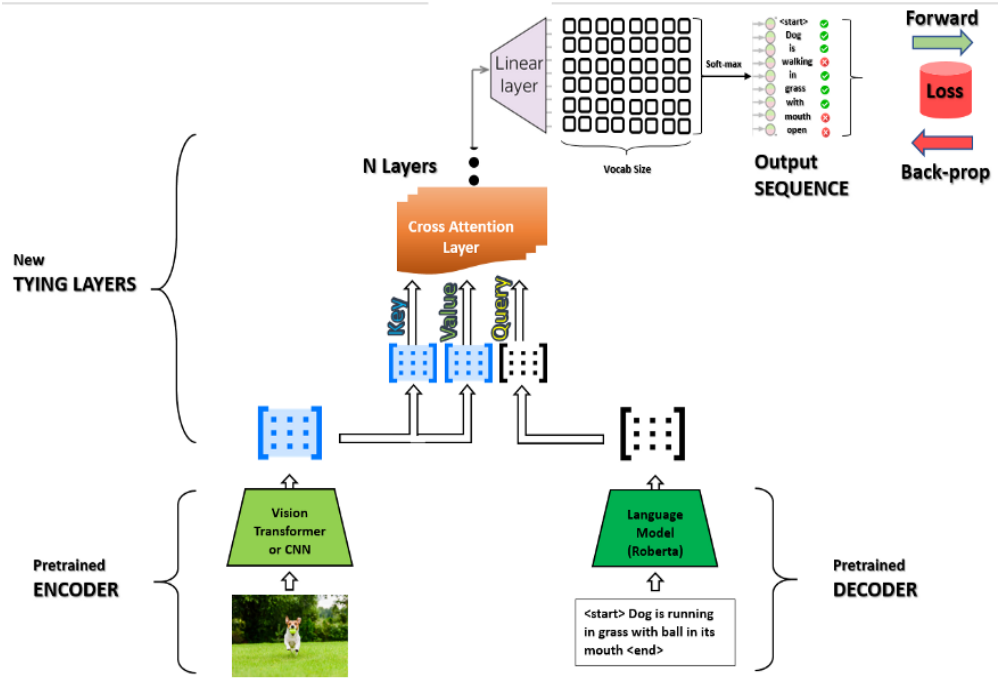
In the decoder part, of the encoder-decoder architecture for image captioning, we used a Language Model which is also based on the Transformers family. We performed several experimentations invoking different Pretrained Language Models such as Bert, GPT2 and Roberta. Every model has its own tokenizer, therefore there are main differences in the special tokens map. We should mention also that the training of the language model may differ because Bert and Robert are Masked Language Model, whereas the GPT2 is Generative Language model.

Our main goal of the Language model is to finetune the respective model with our corpus of text, which will allow the decoder to learn new words and generate brief captions. This will also fine tune the self-attention weights of the transformer to build better context of sentences.

A very important step in the encoder-decoder architecture, is the connection between the encoder and the decoder. This can be achieved **through the cross-attention layer**, in which we utilize and apply the attention mechanism from the output of the encoder as key and value and the output of the decoder as query. In this way, we map the Encoder output with the Decoder input, allowing the Decoder to decide which Encoder input is relevant to put focus on.

Finally, after the cross-attention layer we can generate a brief caption of image with respect of the encoder and decoder. In this way, image captioning is analogous to other sequence-to-sequence tasks except the only difference is instead of translating from a language to language, such as the original Transformer architecture does, it translates from an Image to a language

Fig. 4.4 A high level architecture of the Encoder-Decoder along with the cross-attention layer

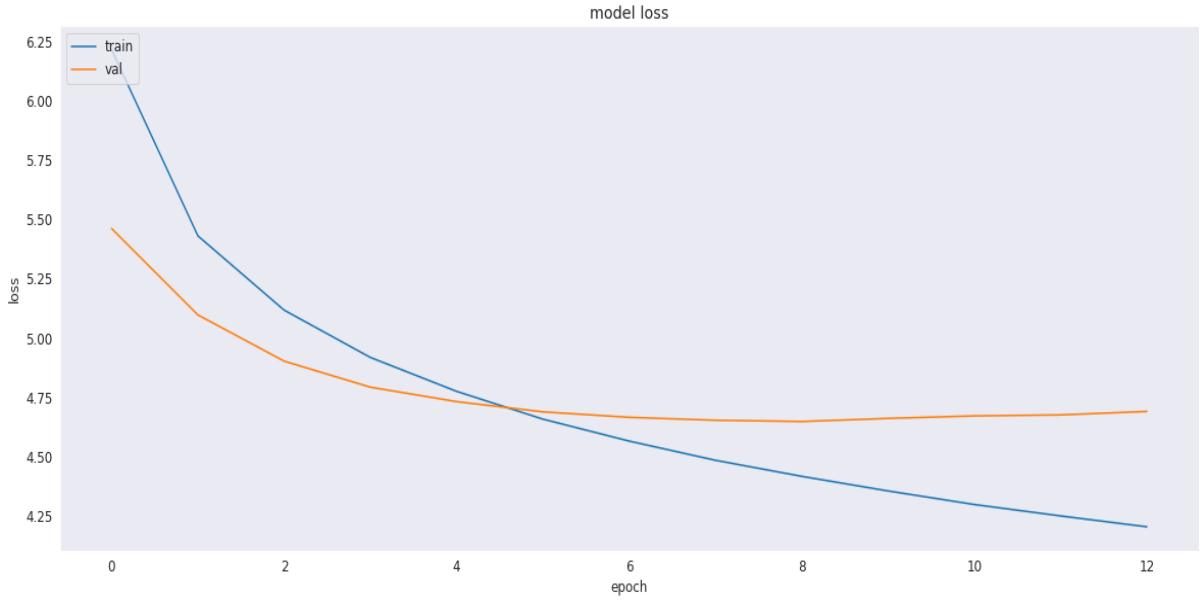


5.Models

5.1 DenseNet & LSTM

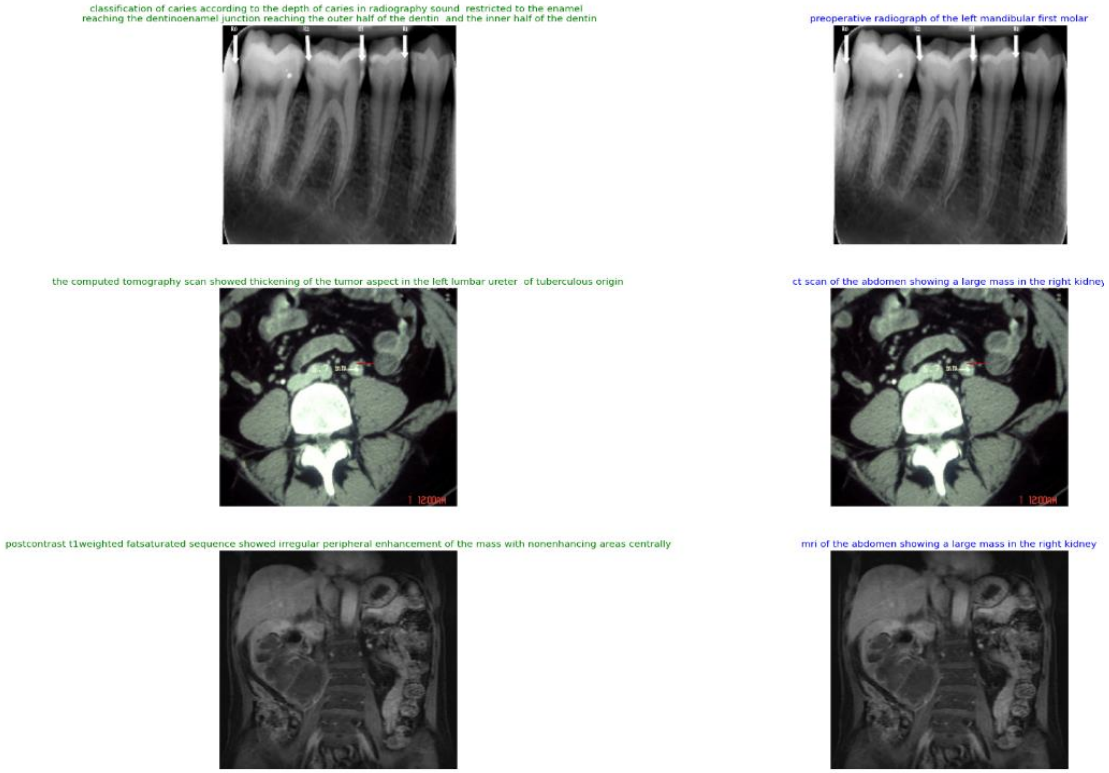
With the conventional architecture of Convolutional Neural Networks and LSTMs, we concatenated the word embeddings and the vector embeddings as we elaborated above and with respect to the pretrained Dense Net201, we trained the model for 20 epochs. We should mention that we saved the “best” model with respect to the validation loss and applied Early Stopping as well. As we can notice from the training graph below, the best validation loss is marked at 8th epoch and the training is stopped at the 13th epoch due to Early Stopping.

Fig. 5.1.1 The training and validation loss of the Dense Net & LSTM model along its epochs



From the generated captions below from the Dense Net & LSTM model, we can notice that the captions are not only brief but also that the model has a tendency to repeat itself. From the images below ***we should mention that the original captions in green and the generated captions in blue***.

Fig. 5.1.2 Sample of generated captions (blue) of the Dense Net & LSTM model, along with the original captions (green)



5.2 ViT & Roberta

For the ViT & Roberta model, we implemented the encoder-decoder architecture. In more detail, we invoked the pretrained ViT model **[4]** as an Encoder and the pretrained Roberta Language Model as a Decoder **[5]**.

The particular ViT model process 224x224 images and splits the respective image into 16 patches. Therefore, we resized the images accordingly before extracting the features of the image. The Roberta base model is a pretrained model on English language using a masked language modelling objective. Along with the pretrained model, we invoked and its respective tokenizer. However, the decoder of the model has to be finetuned in our training corpus, and in this way, we perform a masked modelling in order the decoder to adapt to our target captions. Therefore, we train the Roberta for 10 epochs monitoring not only the training and validation loss per epoch but also their perspective perplexity.

As a last step, we connect the encoder and the decoder through the cross attention and train the model end to end for 5 epochs.

Fig. 5.2.1 A evaluation of the finetuned Masked language model of Roberta (Decoder), from testing corpus

```
masked sentence : <mask> appearance in sonography
unmasked word : Hydronephrosis
predicted word : "normal" with maximum score : 0.6298666596412659

masked sentence : focal steatosis of the <mask> parenchyma
unmasked word : liver
predicted word : " liver" with maximum score : 0.8790934085845947

masked sentence : showing the subtrochanteric fracture in the porotic <mask>
unmasked word : bone
predicted word : " region" with maximum score : 0.18236856162548065

masked sentence : computed tomography <mask> in axial view showing obliteration of the left maxillary sinus
unmasked word : scan
predicted word : " scan" with maximum score : 0.6919050216674805

masked sentence : view of giant cell <mask> of thumb metacarpal preoperatively
unmasked word : tumor
predicted word : " tumor" with maximum score : 0.49643972516059875

masked sentence : <mask> showing high signals involving the superior sagittal sinus thrombosis on TW1
unmasked word : MRI
predicted word : "image" with maximum score : 0.19878479838371277
```

Fig. 5.2.2 The training graph of the Decoder (finetuning the Roberta language model)

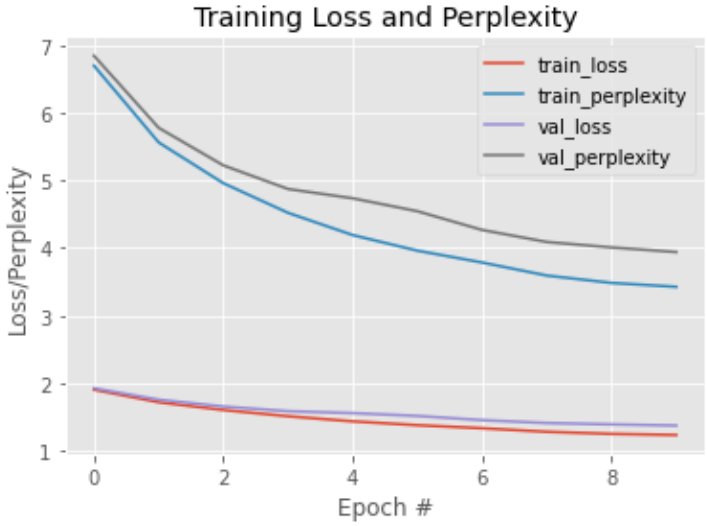


Fig. 5.2.3 The training process of the Encoder and the Decoder (ViT & Roberta)

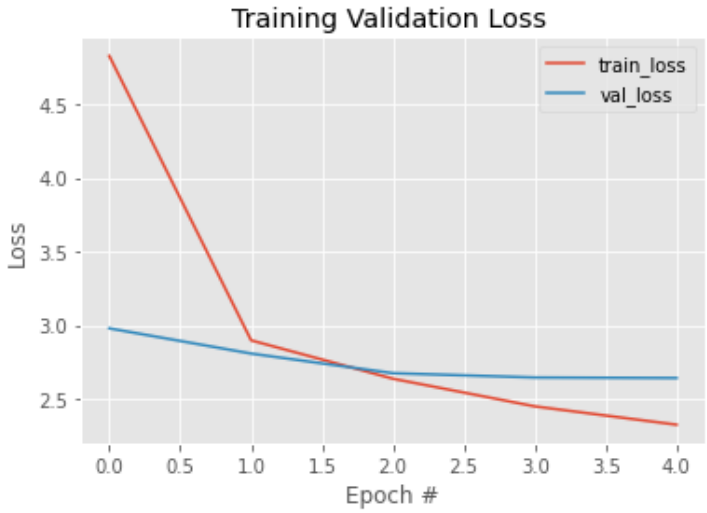
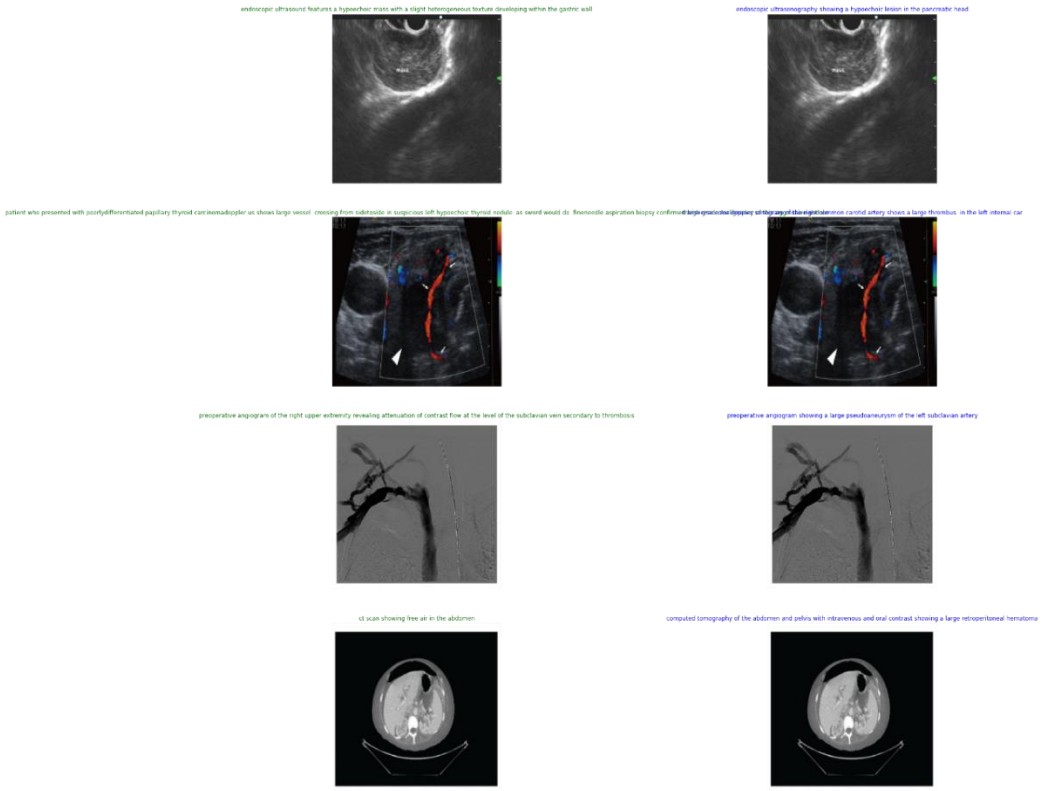


Fig. 5.2.4 Sample of generated captions (blue) of the ViT & Roberta model, along with the original captions (green)



5.3 ViT & GPT2

For the ViT & GPT2 model, we applied the same procedure as the ViT & Roberta, but the only difference is that we invoked the GPT2, generative model **[7]** along with its tokenizer, as a decoder. We trained the decoder to our corpus for 10 epochs, monitoring the training and validation loss and the respective perplexities.

As a last step we tied the Encoder & Decoder into a Vision Encoder-Decoder, and we finally trained the model for 5 epochs. We should mention that although the training loss was decreasing along the epochs, the validation loss was increasing, thus indicating that the model will have a poor performance in the validation/test dataset.

Fig. 5.3.1 The training graph of the Decoder (finetuning the GPT2 language model)

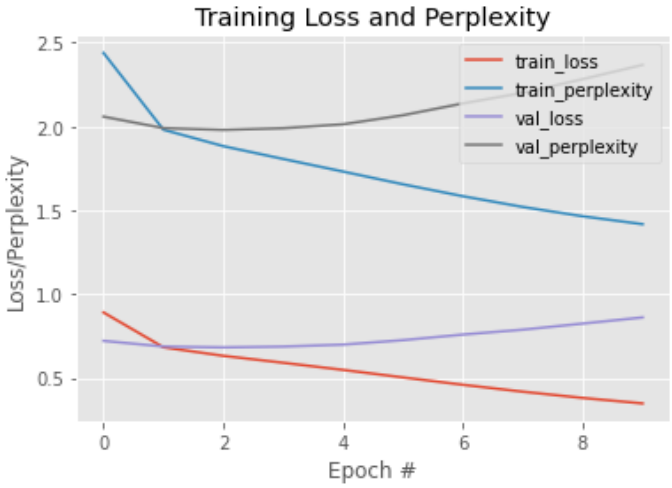
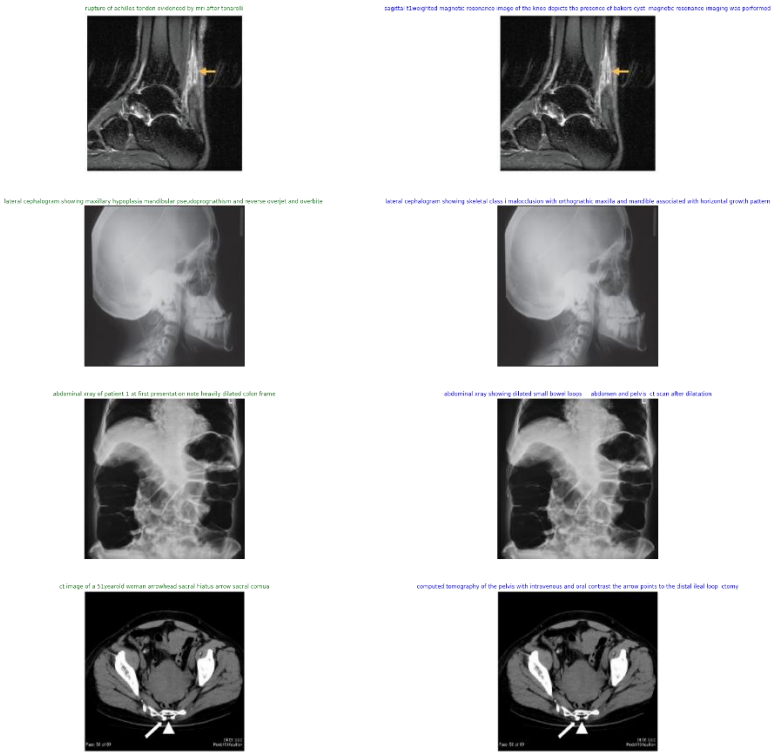


Fig. 5.3.2 A evaluation of the finetuned Generative language model of GPT2 (Decoder), from testing corpus

```
0:  panoramic radiograph obtained after trauma showing a wellcircumscribed uniloculer radiolucent lesion involving right maxilla
1:  panoramic radiograph of patient
0:  the patient underwent radiotherapy and presented a lower left third molar tooth fracture
1:  the patient underwent a folfirino spin echo t2weighted spin echo mr documenting spinous process compression
0:  tumor and didelphys uterus 2nd week of pregnancy
1:  tumor mass of the left parotid gland
0:  liver cystic teratoma
1:  liver cyst on ct scan
0:  ct scan showing the expansion of the mandibular buccal and lingual cortical plates with no periosteal reaction or soft tissue edema
1:  ct scan showing the tumor in the nasal septum
0:  angiography showing a large aneurysm arising from the right basal and anterior circulation of the right lower lobe
1:  angiography of the right kidney showing the presence of the collateral veins in the lower pole of the left kidney
0:  chest xray showing right hilar prominence
1:  chest xray showing leftsided pleural effussion
0:  scan showing a lesion with moderate signal intensity in the center of the left breast
1:  scan showing multiple liver lesions
```

Fig. 5.3.3 The training process of the Encoder and the Decoder (ViT & GPT2)

Fig. 5.3.4 Sample of generated captions (blue) of the ViT & GPT2 model, along with the original captions (green)



5.4 ViT & BioBert

The ViT & Bio Bert model follows the same encoder-decoder approach as the precious mentioned models. However, as a decoder we set the pretrained Bio Clinical Bert model from Hugging face **[7]** along with its tokenizer. As the title of the pretrained model implies the Bio Clinical Bert is based on the Bert architecture, but it was trained on all notes from MIMIC III, a database containing electronic health records from patients at Boston MA hospital. In this way, this particular language model is already familiar with the medical terminology thus **expecting a better performance as a Decoder**. The encoder of the ViT & Bio Bert model is the pre-mentioned ViT model **[4]**.

As a next step we train the Bio Bert decoder for 10 epochs to our ROCO captions and apply Masked Language Modelling. Finally, we can connect the encoder-decoder and train the model end to end for 5 epochs. We should mention that during the training process of the Vision-Encoder Decoder model the validation loss was slightly decreased.

Fig. 5.4.1 The training graph of the Decoder (finetuning the Bio Bert language model)

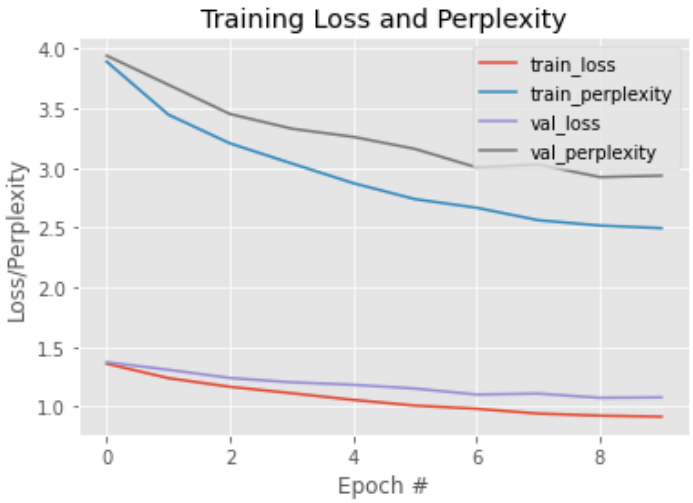
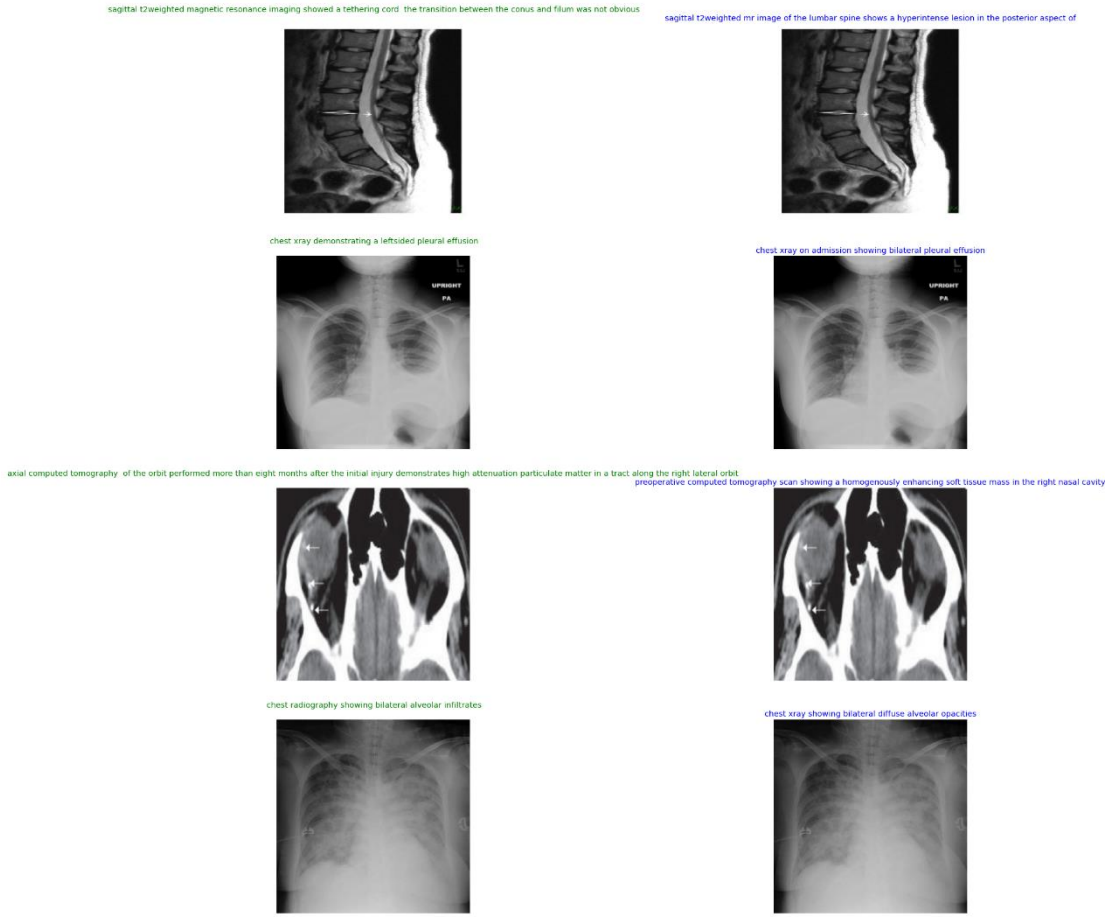


Fig. 5.4.2 A evaluation of the finetuned Masked language model of Bio Bert (Decoder), from testing corpus

```
masked sentence : computed tomography [MASK] in axial view showing obliteration of the left maxillary sinus
unmasked word : scan
predicted word : "scan" with maximum score : 0.866060197353363

masked sentence : view of giant cell [MASK] of thumb metacarpal preoperatively
unmasked word : tumor
predicted word : "tumor" with maximum score : 0.9578973650932312

masked sentence : [MASK] showing high signals involving the superior sagittal sinus thrombosis on TW1
unmasked word : MRI
predicted word : "image" with maximum score : 0.32701238989830017

masked sentence : mdct angiography showing the [MASK] of coa in a 1-month-old girl
unmasked word : location
predicted word : "presence" with maximum score : 0.34742575883865356

masked sentence : chest radiograph obtained after endoscopic submucosal [MASK] showing left pleural fluid with subsegmental collapse
unmasked word : dissection
predicted word : "drainage" with maximum score : 0.7763930559158325

masked sentence : case 2 tibial [MASK]
unmasked word : fracture
predicted word : "fracture" with maximum score : 0.8791098594665527
```

Fig. 5.4.3 The training process of the Encoder and the Decoder (ViT & Bio Bert)

Fig. 5.4.4 Sample of generated captions (blue) of the ViT & Bio Bert model, along with the original captions (green)



5.5 Experiment with Greek Bert

In this section, we tried to implement a Biomedical Image captioning but with Greek captions. Because the ROCO dataset **[3]** have English captions we invoked from Hugging face a pretrained translation model **[9]**, in order to have the respective Greek captions.

Unfortunately**, the transition from English to Greek led to some translation errors** as we can see from the figure below and we believe that this is mainly due to the fact that the respective translation model was not trained on the English-Greek medical terminology.

Fig. 5.5.1 Sample from original captions of ROCO in contrast to the Greek translated captions

captions	captions_gr
Computed tomography scan in axial view showing obliteration of the left maxillary sinus	υπολογισμενη τομογραφια σε αξονικη αποψη που δειχνει εξαλειψη του αριστερου μειστου αμινου
Bacterial contamination occurred after completion of root canal treatment in the tooth, which remained with a temporary filling for 15 month.	η βακτηριακη μολυνση πραγματοποιηθηκε μετα την ολοκληρωση της ριζικης διωρυγας στο δοντι η οποια παρεμεινε με προσωρινη πληρωση για 15 μηνες
The patient had residual paralysis of the hand after poliomyelitis. It was necessary to stabilize the thumb with reference to the index finger. This was accomplished by placing a graft from the bone bank between the first and second metacarpals. The roentgenogram shows the complete healing of the graft one year later.	ο ασθενης ειχε υπολειμματικη παραλυση του χεριου μετα την πολιομυελιτιδα ηταν απαραιτητο να σταθεροποιηθει ο αντιχειρας με αναφορα στο δαχτυλο του δεικτη αυτο επιτευχθηκε με την τοποθετηση σταφυλου απο την τραπεζα οστων μεταξυ του πρωτου και του δευτερου μετακαρδου το ροεντζενογραμμα δειχνει την πληρη θεραπεια του μοσχους ενα χρονο αργοτερα
Panoramic radiograph after immediate loading.	πανοραμικη ακτινογραφια μετα την αμεση φορτωση
Plain abdomen x-ray: Multiple air levels at the mid-abdomen (arrows), no radiopaque shadow, and no air under the diaphragm.	απλη κοιλιακη χωρα χ ray πολλαπλα επιπεδα αερα στη μεση της κοιλιακης περιοχης χωρις ραδιοπακτινοπακτικη σκια και χωρις αερα κατω απο το διαφραγμα
A 3-year-old child with visual difficulties. Axial FLAIR image show a supra-sellar lesion extending to the temporal lobes along the optic tracts (arrows) with moderate mass effect, compatible with optic glioma. FLAIR hyperintensity is also noted in the left mesencephalon from additional tumoral involvement	ενα τριτο παιδι που αντιμετωπιζει οπτικες δυσκολιες axial flair εικονας δειχνει μια υπερενταση υπερεντασης supra η οποια επεκτεινεται στους κροταφικους λοβους κατα μηκος των οπτικων ιχνων με μετρια μαζικη επιδραση συμβιβασιμη με τον οπτικο glioma flair επισημαινεται επισης στην αριστερη mesencephalon απο προσθετη σωματικη συμμετοχη
Showing the subtrochanteric fracture in the porotic bone.	δειχνει το υποτροπικο καταγμα στο porotic οστο
Post orthodontic treatment. Root canal therapy done with maxillary incisors	ταχυδρομικη ορθοδοντικη θεραπεια ρωμαικη θεραπεια καναλιων που πραγματοποιειται με μεγιστους ενοχλητικους
Two sequential thrombi in the distal segment of the obtuse marginal 2 (OM2).	δυο διαδοχικες θρομβιες στο μακρινο τμημα του περιθωριου obtuse 2
An example of MRI image that takes advantage of joint effusion as contrast material in acute scenario. 57x46mm (150 x 150 DPI).	παραδειγμα εικονας μαγνητικης που εκμεταλλευεται την κοινη αποδοση ως αντιθετως υλικο σε οξυ σεναριο 57x46mm

However, we tried to experiment with those translation along with the Greek version of Bert **[8].** A necessary pre-processing action for the Greek captions is to remove the accents from the words. We re-execute the visualization of a word cloud so that can give us a better understanding of the most frequent terms.

The steps remain the same such as training of the Decoder to our Greek corpus, before connecting the decoder to the pretrained ViT encoder **[4].**

From the below training figures of the Vision-Encoder Decoder model, we can notice that the validation loss was slightly decreased, but the generated Greek captions from the ViT & Greek Bert model can be considered as a good estimation.

Although the assumptions that we made as the Greek translated captions had an integrated error, the overall performance of the model is acceptable.

Fig. 5.5.2 A word cloud of the Greek translated captions

Fig. 5.5.3 A evaluation of the finetuned Masked language model of Greek Bert (Decoder), from testing corpus

```
masked sentence : υπολογισμενη [MASK] σε αξονικη αποψη που δειχνει εξαλειψη του αριστερου μεγιστου αμινου
unmasked word : τομογραφια
predicted word : "τομογραφια" with maximum score : 0.9996954202651978

masked sentence : η σπονδυλικη [MASK] ct δειχνει μαζα σε επιπεδο c3 με σημαδεμενη καταστροφη οσων
unmasked word : στηλη
predicted word : "στηλη" with maximum score : 0.9178794622421265

masked sentence : μεταθανατια πανοραμικη ακτινογραφια μετα το [MASK] το κοστοχονδρικο μοσχευμα προσαρμοσμενο στην αριστερη περιοχη
unmasked word : χειρουργειο
predicted word : "τραυμα" with maximum score : 0.346522718667984

masked sentence : στην αριστερη μανδιβουαλικη ramus κατα πεντε ποδηλατικες [MASK] για ακαμπτη εσωτερικη στερεωση
unmasked word : βιδες
predicted word : "προβολες" with maximum score : 0.12536652386188507

masked sentence : αλιευματα των [MASK] με νημα ραμματος 50 που ραμματος και την ανατομη τους κοντα στην εισαγωγη προσθεση ματιων απ
unmasked word : μυων
predicted word : "ματιων" with maximum score : 0.9416866898536682

masked sentence : μεταθανατια πλευρικη [MASK] ενος γυναικειου ασθενη τρεις μηνες μετα την εγχειρηση
unmasked word : μαγνητικη
predicted word : "ακτινογραφια" with maximum score : 0.5733089447021484

masked sentence : παραπλευρη κυκλοφορια απο την [MASK] του κωνου στην αριστερη προσθια αρτηρια που κατεβαινει
unmasked word : αρτηρια
predicted word : "αρτηρια" with maximum score : 0.16663919389247894
```

Fig. 5.5.4 The training graph of the Decoder (finetuning the Greek Bert language model)

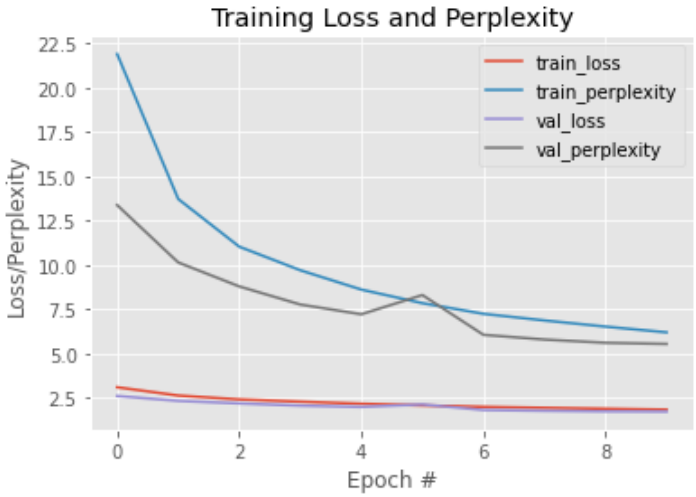
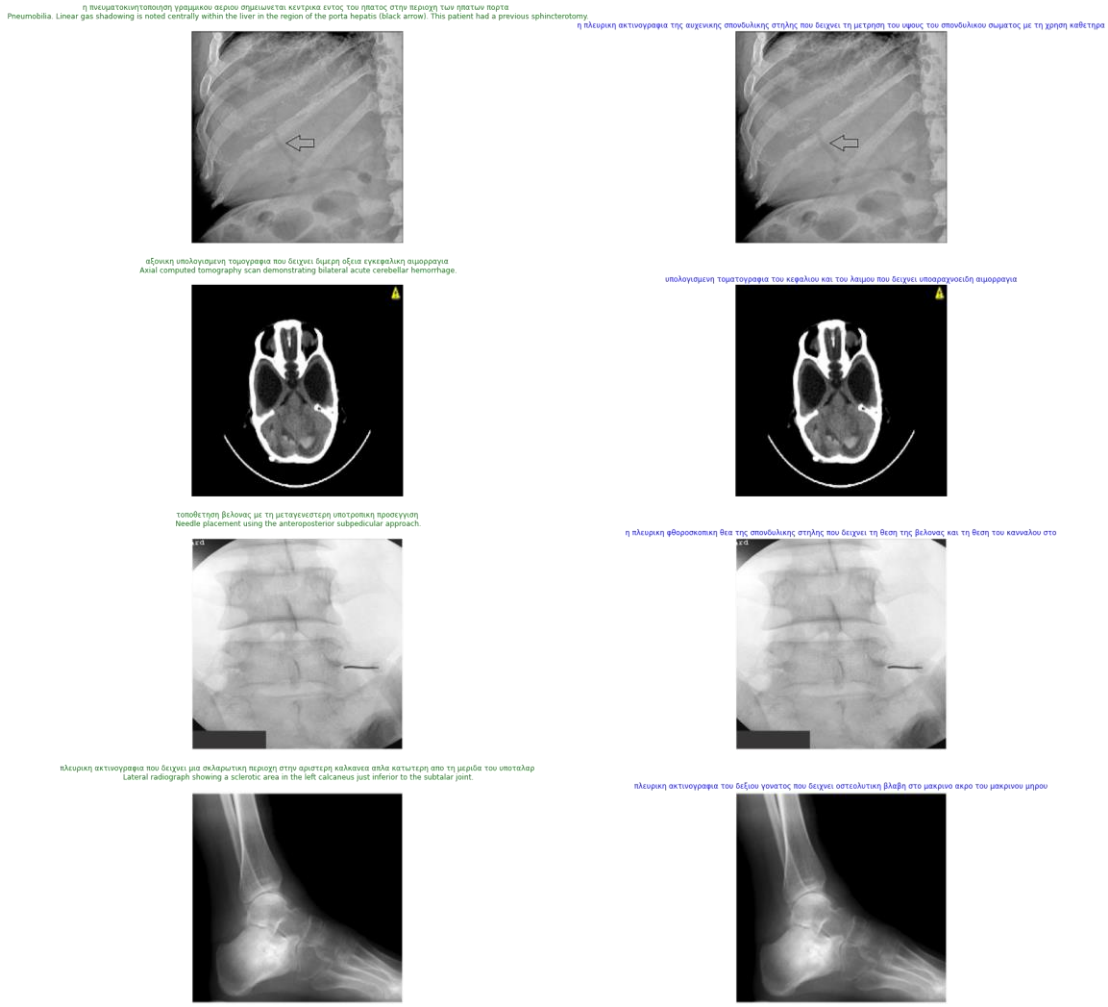


Fig. 5.5.5 The training process of the Encoder and the Decoder (ViT & Greek Bert)

Fig. 5.5.6 Sample of generated captions (blue) of the ViT & Greek Bert model, along with the original captions (green)

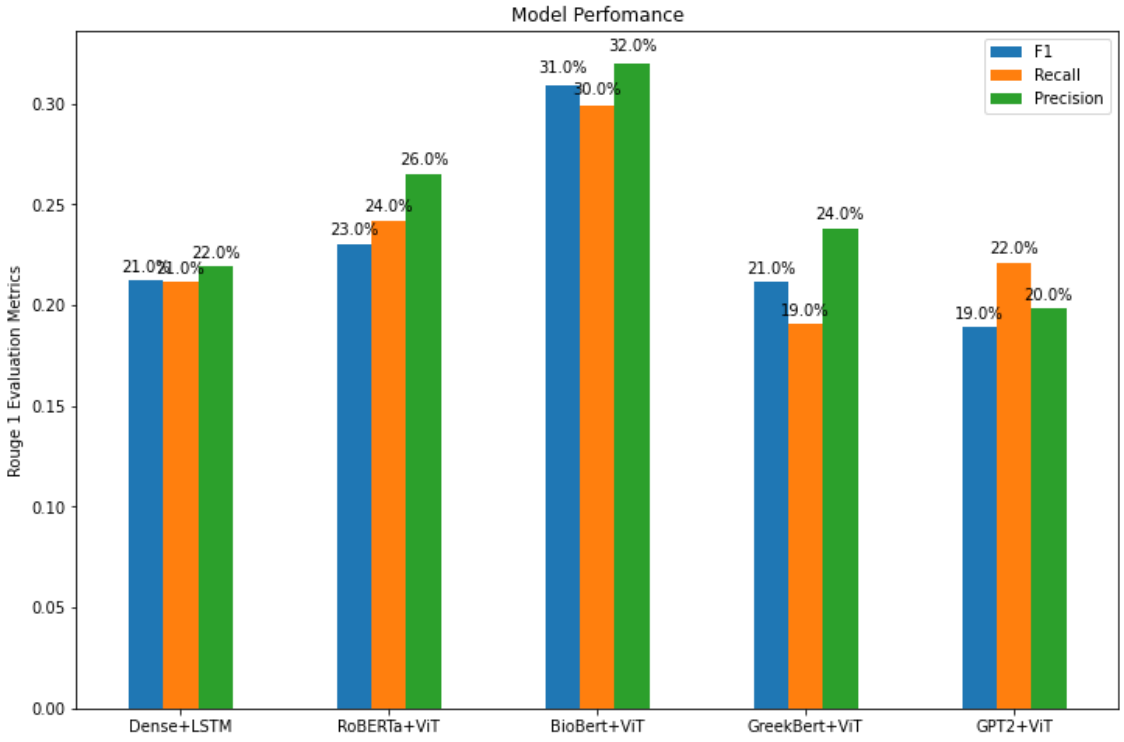


6. Overall Performance

In order to evaluate the overall performance of all the above models, we took a subset of the validation set and applied the Rouge-1 evaluation NLP metric, which measures the quality of the generated caption with respect the original caption. Essentially, the Rouge-N evaluation metric compares the n-grams of the generated summary to the n-grams of the original summary. In our case, we utilized the Rouge-1 metric which focuses only to unigrams. In this way, we computed the Rouge-1 Precision, Recall and their respective harmonic men, the Rouge-1 F1 evaluation metric.

As we can notice from the bar chart of the Model Performance below, **the encoder-decoder model of ViT & Bio Bert scored the greatest Rouge-1 F1 metric with 31 %**. The encoder-decoder ViT & Roberta model marked a Rouge-1 F1 metric of 23 %, which is significantly less than the ViT & Bio Bert model. The models of Vit& Greek Bert and the Dense Net & LSTM achieved a similar Rouge-1 F1 metric of 21%. Finally, we should mention that the ViT & GPT2 model marked the lowest performance in terms of Rouge-1 F1 evaluation metric, achieving a Rouge-1 F1 score of 19%, which is even lower than the conventional Dense Net & LSTM model of 21%.

Fig. 6 A visualization of all the models with respect to the Rouge1 F1, Precision and Recall metric



7. Conclusion

To sum up, in this paper we explored the different aspects of the Image Captioning task. We focused to generate captions mainly from Biomedical X ray images. Due to the nature of this task, we mentioned that is very challenging not only to evaluate the generated captions with the original captions, but also to interpret the biomedical generated captions in contrast to the original captions.

In conclusion, we compared and implemented two different architectures of Image captioning, the conventional approach of CNN and LSTM and the contemporary approach of Encoder-Decoder along with attention. Finally, with respect to the evaluation metrics and the overall performance of the models, we noticed that the **encoder-decoder models are more robust to image captioning task, outperforming the older architecture models**.

References

[1] "**Attention is all you need**" by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan Gomez, Lukasz Kaiser, Illia Polosukhin (arxiv)

[2] "**An image is worth 16x16 words, Transformers for Image Recognition at Scale**", By Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn , Xiaohua Zhai, Thomas Unterthiner , Mostafa Dehghani ,Matthias Minderer, Georg Heigold , Sylvain Gelly, Jakob Uszkoreit , Neil Houlsby (arxiv)

[3] **ROCO** source dataset (link), (kaggle link)

[4] **ViT** model from Hugging Face (google/vit-base-patch16-224), (link)

[5] **Roberta** model from Hugging Face (roberta-base), (link)

[6] **BioBert** model from Hugging Face (emilyalsentzer/Bio_ClinicalBERT) , (link)

[7] **GPT2** model from Hugging Face (gpt2) , (link)

[8] **Greek Bert** model from Hugging Face (nlpaueb/bert-base-greek-uncased-v1) , (link)

[9] Hugging Face **translation model** (Helsinki-NLP/opus-mt-en-grk) , (link)