



AUEB NLP Group at ImageCLEFmedical Caption 2023

P. Kaliosis¹, G. Moschovis^{1,2}, F. Charalampakos¹,
J. Pavlopoulos¹, I. Androutsopoulos^{1,2}

¹ Department of Informatics, Athens University of Economics and Business, Greece

² Archimedes Research Unit, Athena Research Center, Athens, Greece

You can find out more on our work here: nlp.cs.aueb.gr/



Scan for slides



Outline

- Diagnostic Captioning ←
- Task #1: Concept Detection
 - Task Breakdown
 - Methods
- Task #2: Caption Prediction
 - Task Breakdown
 - Methods
- Results & Future Work

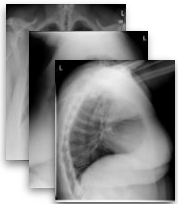


Diagnostic Captioning

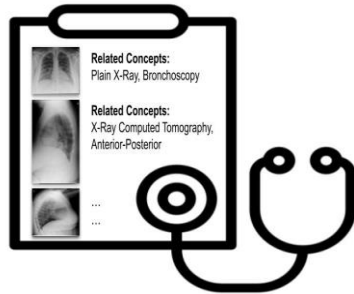
- **Challenging** research problem → Helpful only as an **assistive** tool, not a replacement of the medical staff.
- **More experienced** clinicians → **improve** throughput and accuracy.
- **Less experienced** clinicians → **consider** the generated medical report & reduce the possibility of a clinical error.

Suggested DL-based pipeline for the diagnosis of a given medical image:

Medical Images



Generated report from DL system



Doctor considers the generated report



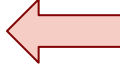
And composes the final report!





Outline



- Diagnostic Captioning
- **Task #1: Concept Detection** 
 - Task Breakdown
 - Methods
- Task #2: Caption Prediction
 - Task Breakdown
 - Methods
- Results & Future Work



Task #1: Concept Detection



Goal: given a radiology image; predict relevant biomedical concepts → **multi-class, multi-label classification** task.



CC BY [Khougali et al. (2021)]

Desired answer from our Diagnostic Captioning deep learning system: **C0041618;C0238207;C0030797;C0022646;C0006736;C0549186**

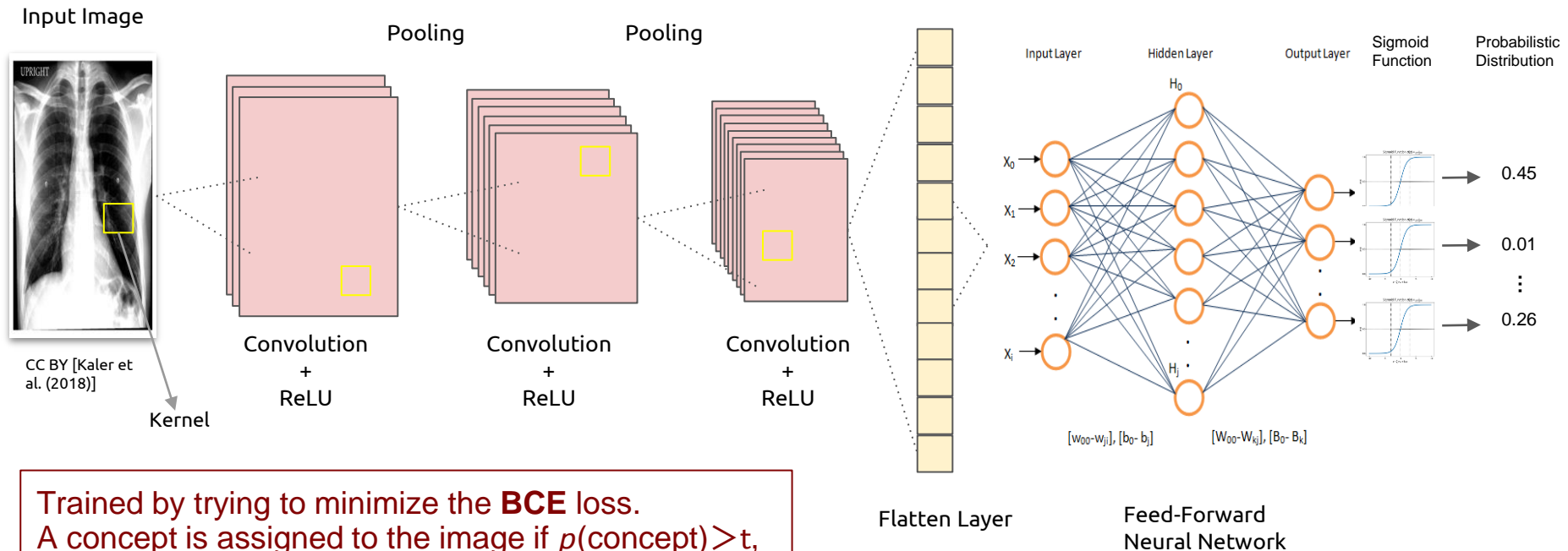


This image is associated with the following biomedical concepts:
Ultrasonography, Ectopic kidney, Pelvis, Kidney, Calculi, Obstructed



System #1: CNN + FFNN

- **CNN**: in charge of encoding the image into numeric vectors. **FFNN**: responsible for yielding a probability distribution over the medical concepts.
- Simple, yet effective idea.

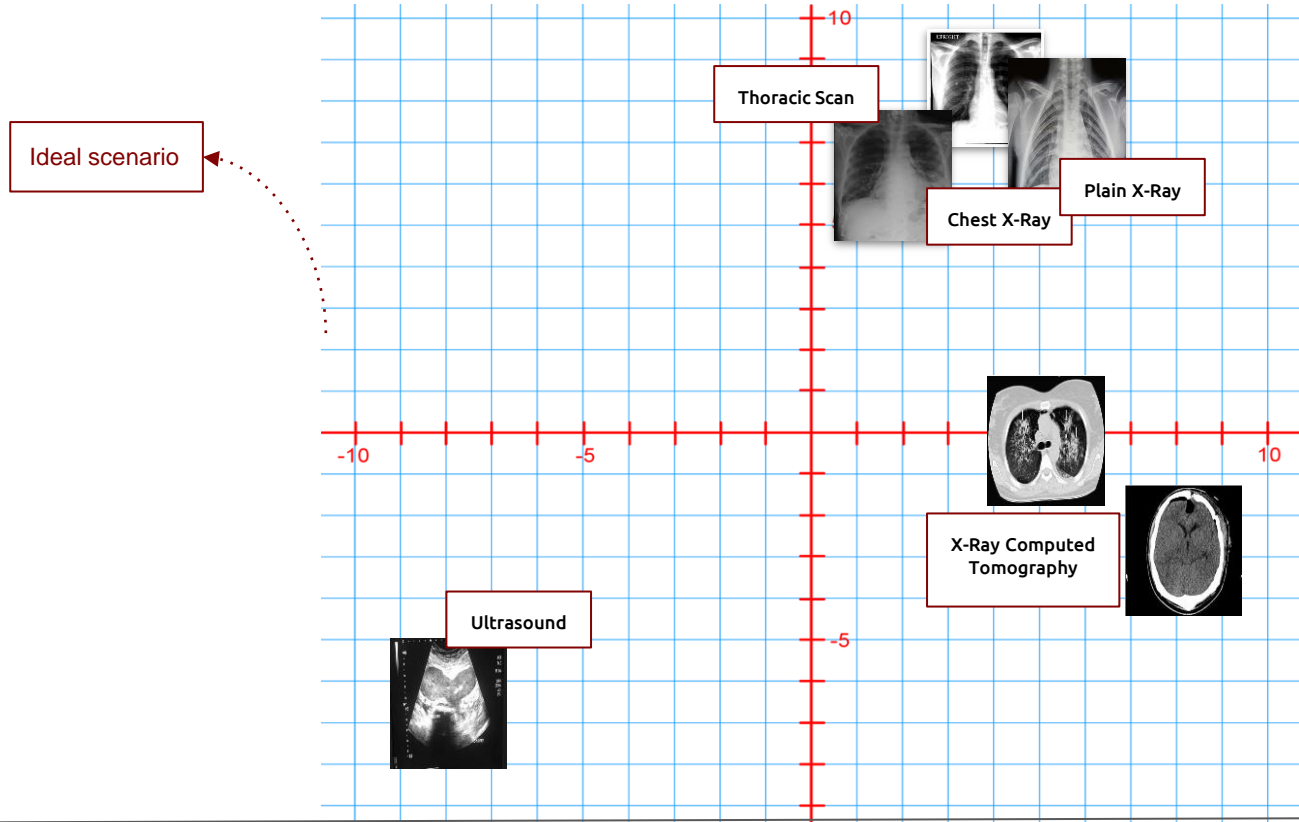


Trained by trying to minimize the **BCE** loss.
A concept is assigned to the image if $p(\text{concept}) > t$,
where **t** is a tunable **threshold** value.



System #2: Contrastive Learning-based Tagger

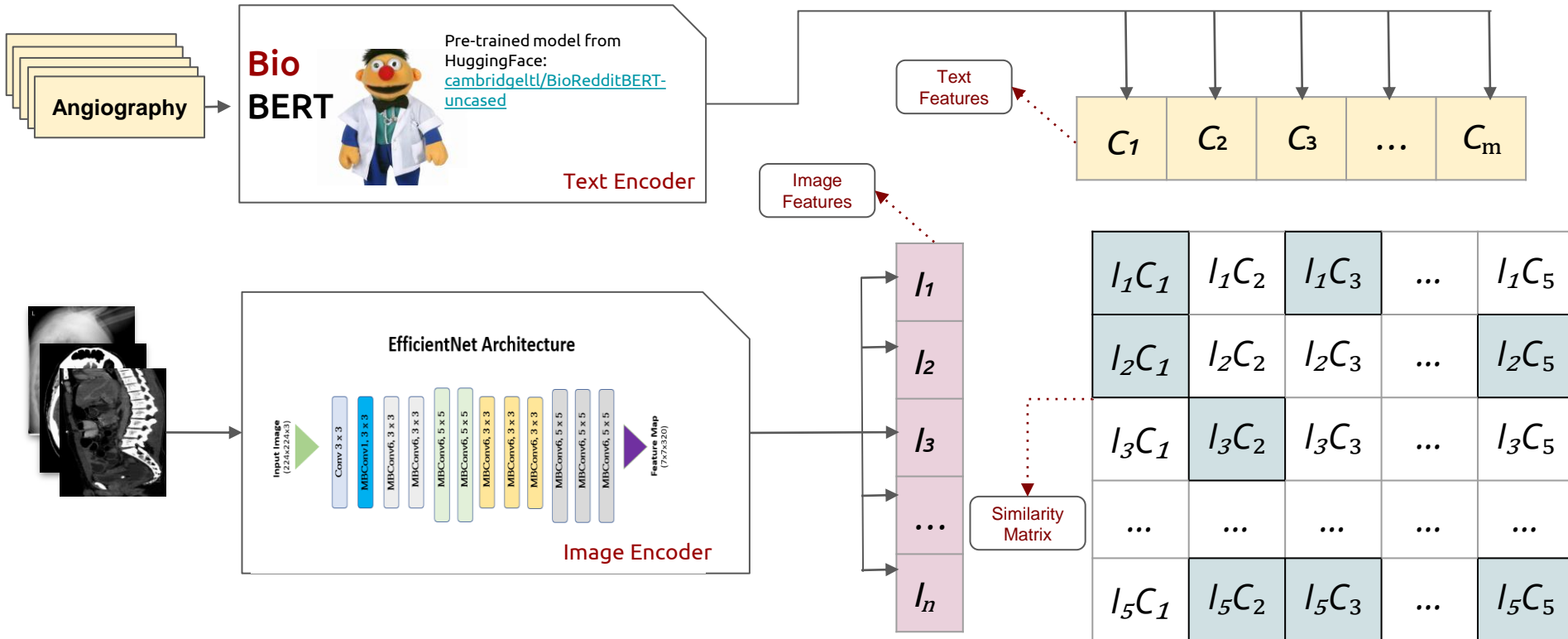
Goal: Train a multi-modal model w.r.t a contrastive objective; bring representations of true pairings closer in the vector space, while pushing the representations of mismatching pairs far away.





System #2: Contrastive Learning-based Tagger

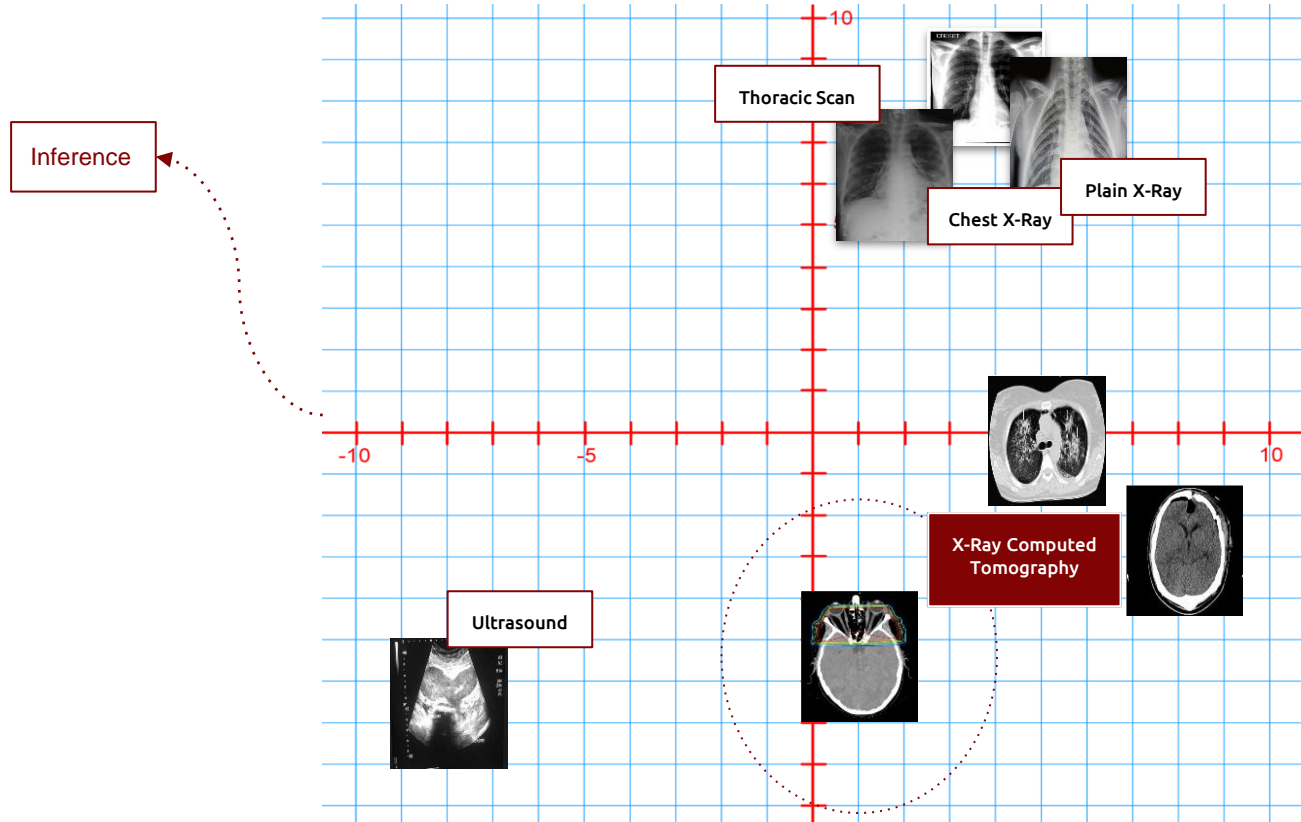
Goal: Train a multi-modal model w.r.t a contrastive objective; bring representations of true pairings closer in the vector space, while pushing the representations of mismatching pairs far away.





System #2: Contrastive Learning-based Tagger


Goal: Train a multi-modal model w.r.t a contrastive objective; bring representations of true pairings closer in the vector space, while pushing the representations of mismatching pairs far away.





Outline



- Diagnostic Captioning
- Task #1: Concept Detection
 - Task Breakdown
 - Methods
- **Task #2: Caption Prediction** 
 - Task Breakdown
 - Methods
- Results & Future Work



Task #1: Caption Prediction



Goal: given a radiology image; generate a draft diagnostic report → **open-ended** generation task.



Generate
caption...



Chest X-ray showing
bilateral clavicular
hypoplasia.



Generate
caption...

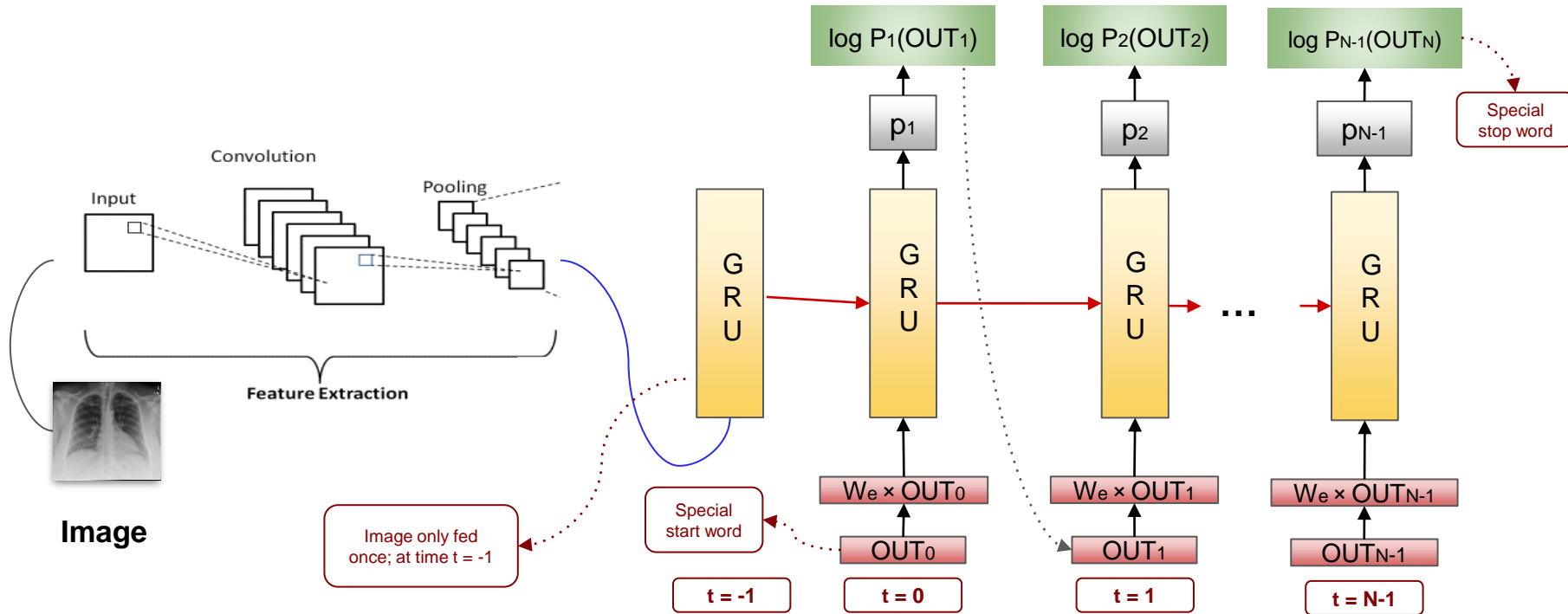


Abdominopelvic ultrasound scan showed
ectopic kidneys at the hemi-pelvis, fused in
their upper poles, normal size and texture of
the kidneys with normal corticomedullary
differentiation, no stones or obstructive
changes.



System #1: CNN-RNN (Show&Tell)

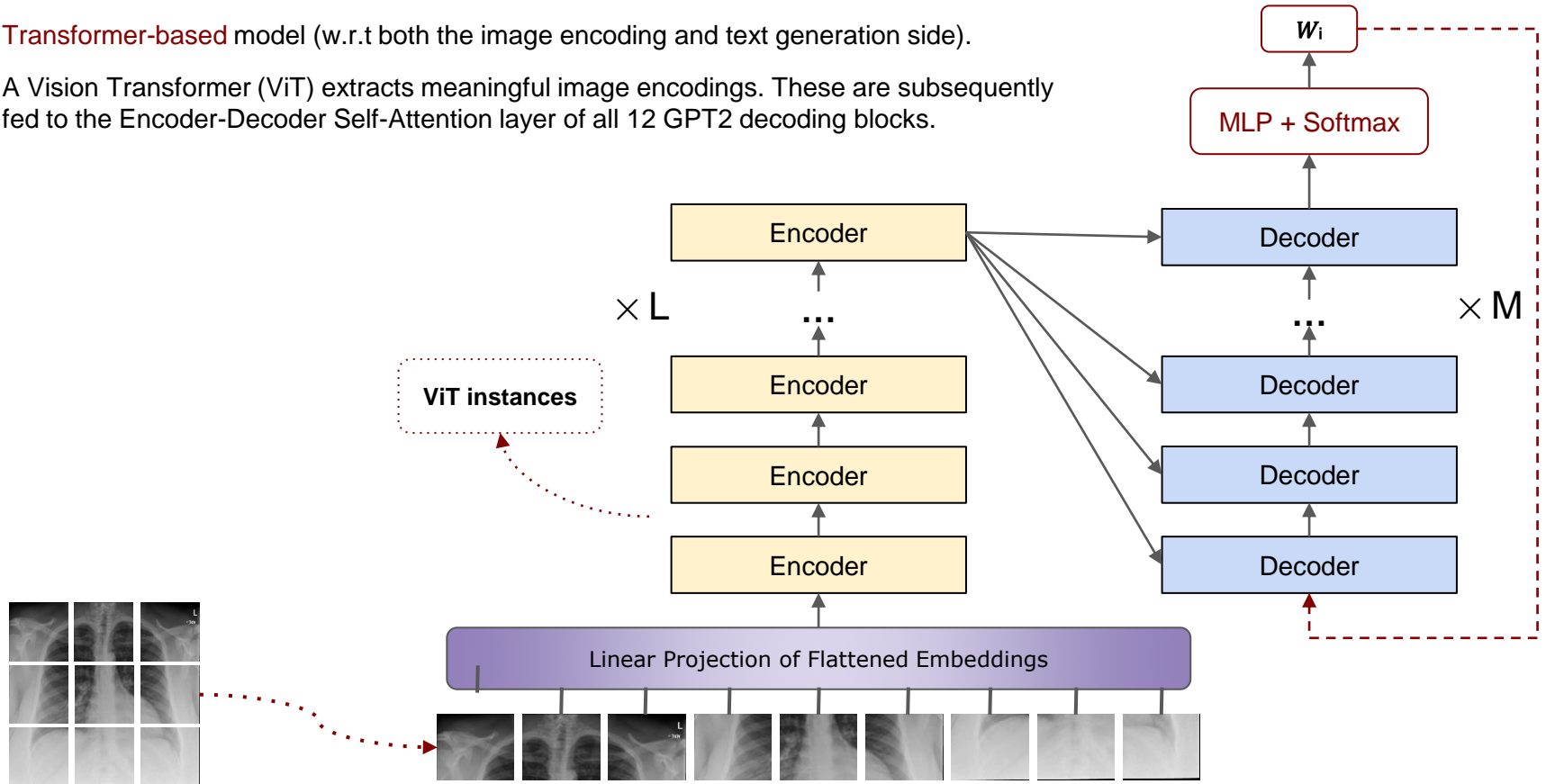
- Well-known **Image Captioning** architecture – **SotA** before the introduction of Transformers.
- **CNN**: in charge of encoding the image and extracting visual features. **RNN**: generates the diagnostic caption based on the visual features.





System #2: ViT-GPT2

- **Transformer-based** model (w.r.t both the image encoding and text generation side).
- A Vision Transformer (ViT) extracts meaningful image encodings. These are subsequently fed to the Encoder-Decoder Self-Attention layer of all 12 GPT2 decoding blocks.

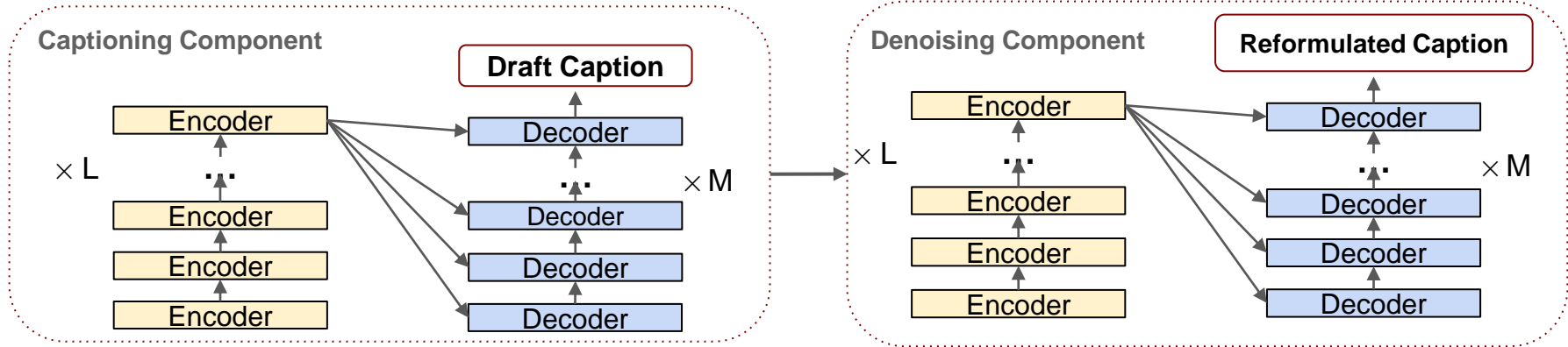




System #3: 2xE-D: Captioning Model + Seq2Seq denoiser



- **2xE-D**: two **Encoder-Decoder** architectures running in **sequential** order.



Experimented with two pre-trained denoising models, BART and T5. BART is a denoising autoencoder model. It is trained by reconstructing text that has been distorted by an arbitrary noise function.

We **fine-tuned** BART towards the goal of correcting our model's common **grammatical**, **syntactical** and also **diagnostic** mistakes.

By feeding it (draft generated caption, ground truth caption) **pairs**, enabling it to learn **common** mistake **patterns** in the corpus generated from the initial captioning model.

How?

Alternative Idea



ClinicalBAR
T
(or ClefBART)




Further **pre-trained** a **BART-large** instance on a **held-out** set of the ImageCLEFmedical dataset → following the **original text corruption** processes.

Then applied as a **denoising** component to the draft diagnostic captions derived from our **captioning** systems.



Outline



- Diagnostic Captioning
- Task #1: Concept Detection
 - Task Breakdown
 - Methods
- Task #2: Caption Prediction
 - Task Breakdown
 - Methods
- Results & Future Work 



Results & Future Work



- In this year's ImageCLEFmedical competition (Caption Task), the **AUEB NLP Group** ranked **1st** in the **Concept Detection** and **3rd** in the **Caption Prediction** sub-tasks, according to the competition's primary evaluation metrics.

- Our best performing Concept Detection system was an ensemble▶ system consisted of three CNN+FFNN ([Concept Detection System #1](#)) instances.
- Our best performing Caption Prediction system was a BART ([Caption Prediction System #3](#)) instance applied on top of our CNN-RNN ([Caption Prediction System #1](#)) captioning component, namely **BART@CNN-RNN**.

Ensemble: combination (Union or Intersection) of two or more instances predictions'.

- Future Work:

- Explore instruction-tuning techniques in conjunction with state-of-the-art LLMs in order to enhance the accuracy and interpretability of the generated diagnostic captions.
- Experiment with few-shot, as well as prominent in-context learning techniques, as restricted data is a common problem, especially in the biomedical domain.



Scan for our paper

Thanks for attending!
Any questions?

You can find out more on our work here: nlp.cs.aueb.gr/



Scan for our paper

Additional optional slides





(Tagging) System #2: CNN+FFNN-based Multi-task Classifier



Motivation: There are **four** main **medical modalities** in the dataset: X-Ray, Computed Tomography, MRI and Ultrasonography which almost never occur concurrently.

