



Aprendizaje profundo para generación de reportes radiológicos: modelos de visión, de lenguaje y multimodales para imágenes médicas

> Denis Parra CS Departament, PUC Chile (iHealth, CENIA, IMFD) Invited talk - Magíster en Informática UACh 2023

This research in collaboration with:

- Cecilia Besa (UC School of Medicine)
- Álvaro Soto (CENIA & UC)
- José Cañete (former CENIA, now Wallmart)
- René Vidal (UPenn)
- Jocelyn Dunstan (UC School of Engineering)
- UC CS Students: Pablo Pino, Pablo Messina
- MSc UC Students: Greg Schuit, Francisco Madariaga
- Undergraduate UC students: Valeria Salas, Itan Felszentein



Pontificia Universidad Católica de Chile





Millennium Institute for Intelligent Healthcare Engineering (iHEALTH) 2021-...

C. Prieto (Director), **M. Andía** (Deputy Director), C Besa, R Botnar, S Chabert, P Estévez, D Mery, D Parra, M Rodríguez, R Salas, S Uribe

Challenges need to be tackled in an interdisciplinary manner



UFMG – DCC Week





CENIA: Centro Nacional de Investigación en Inteligencia Artificial 2021 - ...

A. Soto (Director), M. Peña (Deputy Director), et al.

Principal researchers

- Alvaro Soto, UC
- Bárbara Poblete, UChile
- Eduardo Cerpa, UC
- Paula Aguirre, UC
- Pedro Maldonado, UChile
- Marcela Peña, UC



- Cristóbal Rojas, UC
- Pablo Barceló, UC
- Marcelo Mendoza, USM
- Denis Parra, UC
- Domingo Mery, UC
- + 23 investigadores asociados/jóvenes/colaboradores

Line 1: Deep Learning for Vision and Language

New theories and methods to continue unraveling the potential of Deep Learning (DL) for creating advanced cognitive systems with a focus on vision and language.

Applications to generate new theories about the learning mechanisms behind DL and its generalization abilities. (1) Contrastive pre-training



https://github.com/openai/CLIP

Leaders: Mery, Parra y Soto

Back to Medical Imaging: TOC

- Motivation
- The MIRG task
- Deep Learning
- Vision Models
- Language Models
- Multimodal Models
- Discussion



 Eric Topol's "Deep Medicine" (2019) book indicates that in the US, by 2016, there were 800 million medical scans a year, accounting for about 60 billion images. Scaling this up based only on human labor is challenging



Using Artificial Intelligence (AI) for medical imaging and reporting could help hospitals deal with this large and growing demand

٠



 Eric Topol's "Deep Medicine" (2019) book indicates that in the US, by 2016, there were 800 million medical scans a year, accounting for about 60 billion images. Scaling this up based only on human labor is challenging



• By 1975, the mean time allotted for office visits was 60 min (new patients) and 30 min (returning). By 2019, this had changed to 12 min and 7 min, respectively

 Using Artificial Intelligence (AI) for medical image report generation (MIRG) could help hospitals deal with this large and growing demand AI does not mean replacing radiologists, but rather helping them being more effective and <u>spending</u> <u>more time caring for the patients</u>

The patient-physician interaction



Alwin Yaoxian Zhang, et al.. 2019. Explainable AI: Classification of MRI Brain Scans Orders for Quality Improvement. In Proceedings of (BDCAT '19). DOI:https://doi.org/10.1145/3365109.3368791



- MIRG: Medical Image Report Generation
- Given one or more patient's input image(s), generate a text report of the *findings* section of a radiology report



Manual tags: Calcified Granuloma/lung/upper lobe/right Automatic tags: Calcified granuloma

Comparison: Chest radiographs XXXX. Indication: XXXX-year-old male, chest pain. Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.

Impression: No acute cardiopulmonary process.

Example from the IU X-ray dataset, frontal and lateral chest x-rays from a patient, alongside the report and the annotated tags. XXXX is used for anonimization.



iHealth/CENIA on Radiology Report Generation





Messina, Pino et al (2021) A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images. ACM CSUR



Fig. 2. CNN-TRG model for report-generation



Pino et al (2022) Clinically Correct Report Generation from Chest X-Rays Using Templates, MLMI 2021

B. Conditional Report Generator (CRG)



D. CRG with mined topic



Degenerative changes and potentially post-traumatic changes, identified at the left glenohumeral joint. Hypertrophic changes are seen in the spine.

Messina, Cañete et al (2022) Twostage Conditional Chest X-ray Radiology Report Generation. MedNeurIPS 2022.

The MIRG task – A higher level view



Messina, P, Pino, P et al. (2020) "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images"



Deep Learning

Deep Learning (2012 - ...)



Image from https://mc.ai/aisaturdaylagos-the-torch-panther/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **Imagenet classification** with deep convolutional <u>neural networks</u>. NeurIPS 2012.

Deep Learning (2012 - ...)



Image from https://mc.ai/aisaturdaylagos-the-torch-panther/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). **Imagenet classification** with deep convolutional neural networks. NeurIPS 2012.



Progress in machine classification of images: the error rate (%) of the ImageNet competition winner by year.

Deep Learning (2012 - ...) – Representation learning



Deep Learning (2012 - ...) – Representation learning



Different **optimization objectives** show what different parts of a network are looking for.

n layer index
 x,y spatial position
 z channel index
 k class index







Neuron layer_n[x,y,z]

ChannelLalayer_n[:,:,z]l

layer_n[:,:,:]²

https://distill.pub/2017/feature-visualization/

Olah, et al., "Feature Visualization", Distill, 2017.



Traditional MIRG solution: Deep Learning

Encoder + Decoder Architecture



The cardiomediastinal silhouette is within normal limits. **Calcified right lower lobe granuloma**. No focal airspace consolidation. No visualized pneumothorax or large pleural effusion. No acute bony abnormalities.

Parra, Besa (2022) "Multimodal, multitask and transfer learning for deep radiological report generation"



MIRG Architecture example

A CNN visual encoder + a language model decoder (e.g. LSTM)



Messina, P, Pino, P et al. (2020) "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images"

We need to combine architectures and representations!



Messina, P, Pino, P et al. (2020) "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images"



Modelos de Visión

CNNs: Convolutional Neuronal Networks





CNNs: AlexNet, VGG



AlexNet (Krizhevsky et al, 2011) was the first CNN with significant performance improvement in ImageNet challenge

60 million parameters



VGG [5] introduced by the Visual Geometry Group at Oxford U., increased network depth by using smaller convolution filters (3x3)

138 million parameters

ResNet (CNN): State of the art baseline

ResNet introduces **the residual block** to ease the training of networks that are substantially deeper than AlexNet, VGG, and GoogLeNet.

						\sim	~ .							-	
method	top-1 err.	top-5 err.	-		_	6262	$\Delta \Delta$	1616	Y			777			
VGG [40] (ILSVRC'14)	2	8.43 [†]	esidua		64,/2	3 3 3 3	1 23,12	128	,128	256 256 256	256	256	512./2 .512 .512 .512	512 512 0	
GoogLeNet [43] (ILSVRC'14)	-	7.89	ayer r	<u>a</u>	NJ com,	3k3 com 3k3 com 3k3 com 3k3 com	3x3 com 3x3 com	3x3 com 3x3 com 3x3 com 3x3 com	3x3 conv 3x3 conv 3x3 conv 3x3 conv	3k3 conv 3k3 conv 3k3 conv 3k3 conv	3k3 com bill com bill com	3x3 com	 Comu, L Soli comu Soli comu Soli comu Soli comu 	3x3 conv 3x3 conv avg po 4c 100	
VGG [40] (v5)	24.4	7.1	34-1												
PReLU-net [12]	21.59	5.71	c												
BN-inception [16]	21.99	5.81	er plai	a	4,64,12	L/12 mr, 64 mr, 64 mr, 64	ms, 64 ms, 64	mv, 128 mv, 128 mv, 128 mv, 128	nv, 128 nv, 128 nv, 128	mr, 256 mr, 256 mr, 256	m, 256 m, 256 m, 256	mr, 256 mr, 256 mr, 256	r, 512,/2 nv, 512 nv, 512 nv, 512	mv, 512 mv, 512 pool 000	
ResNet-34 B	21.84	5.71	14-layo	2	747 con	900 343 cc 343 cc 343 cc	343 com	343 00	3x3 co 3x3 co 3x3 co 3x3 co 1x3 com	343 00 343 00 343 00 343 00 343 00	MG M	343 00	3x3 com 3x3 com 3x3 co 3x3 co	3x3 co 3x3 co avg fc 1	
ResNet-34 C	21.53	5.60													
ResNet-50	20.74	5.25													
ResNet-101	19.87	4.60	6-19	tonv, 64	01.12 0000,128	ol, /2 m, 256 m, 256 m, 256 m, 256 onv, 256		* * * * * * * * * * * * * *	→ ^{2/} 19	♦ 000%, 512 000%, 512 000%, 512 000%, 512			5,12	4096 4096 4096	
ResNet-152	19.38	4.49	NG	NS NS	843 c	343 (a.	343 (343 (343 (8	363 (363 (363 (363 (2	2 4 4	
ble 4. Error rates (%) of single-model results on the ImageNet				output lite: 224	output lite: 112	output	1	5200: 28	Induce	a 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2		1	output	output size. 1	

validation set (except [†] reported on the test set).

ResNet-50 has aprox. 23 million parameters

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

ResNet for caption prediction

In 2021, the UC Chile Team won the Caption Prediction task of the international Medical ImageCLEF challenge with a ResNet classification backbone

https://www.aicrowd.com/challenges/imageclef-2021-caption-caption-

prediction/leaderboards





Figure 5: Model diagram. Top N = 23 classified words are selected for the caption.

Castro, V., Pino, P., Parra, D., & Lobel, H. (2021). PUC Chile team at Caption Prediction: ResNet visual encoding and caption classification with Parametric ReLU. ImageCLEF 2021 (pp. 1174-1183).



DenseNet modifies ResNet by concatenating feature maps from previous layers.

DenseNet is more difficult to train than ResNet but it promotes reutilization of features, which can be useful in certain visual tasks



Source: https://lightning.ai/docs/pytorch/stable/notebooks/course_UvA-DL/04-inception-resnet-densenet.html

a) X_0 X_1 $Y_1 = X_0 + X_1$ X_2 $Y_2 = X_0 + X_1 + X_2$ X_3 $Y_3 = X_0 + X_1 + X_2 + X_3$ X_1 $Y_1 = X_0 / X_1$ X_2 $Y_2 = X_0 / X_1 / X_2$ X_3 $Y_3 = X_0 / X_1 / X_2 / X_3$

Figure 1. (a) ResNet and (b) DenseNet.

Zhang, C., Benz, P., Argaw, D. M., Lee, S., Kim, J., Rameau, F., ... & Kweon, I. S. (2021). Resnet or densenet? introducing dense shortcuts to resnet. In *Proceedings of WCACV*

DenseNet for image classification and report generation

Our first method for report generation, CNN-TRG based on text templates, used a DenseNet-101 visual encoder backbone.



Pino, P., Parra, D., Besa, C., & Lagos, C. (2021). Clinically correct report generation from chest x-rays using templates. In Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12 (pp. 654-663). Springer International Publishing.

Visual Transformer

Transformers were introduced in NLP, but they made their way to visual tasks as well as to medical imaging



Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 102802.

Visual Transformer: Medical imaging tasks



Fig. 2. A diverse set of application areas of Transformers in medical imaging covered in this survey.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 102802.

Visual Transformer: Medical imaging tasks



Is this really SOTA? IS NOT !

The sad truth: CNNs still competitive

- Results presented in M2 article are Micro-F1 (and results presented are average over the top-5 results)
- Dataset is imbalanced: Macro-F1 is the most appropriate metric

* M2 Macro F-1 = 0.333 vs. * CXR-TRG Macro F-1 = 0.428

Model	Year	NLG METRICS							CE METRICS			
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	Р	R	F1		
TRANSFORMERS (Vaswani et al., 2017)	2017	0.409	0.268	0.191	0.144	0.157	0.318	-	_	_		
MEMORY TRANSFORMER (Chen et al., 2020c)	2020	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276		
CLINICAL TRANSFORMER (Lovelace and Mortazavi, 2020)	2020	0.415	0.272	0.193	0.146	0.159	0.318	0.411	0.475	0.361		
*CMCL (Liu et al., 2022a)	2021	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-		
\mathcal{M}^2 TR. (Nooralahzadeh et al., 2021)	2021	0.361	0.221	0.146	0.101	0.139	0.266	0.324	0.241	0.276		
\mathcal{M}^2 TR. Prog. (Nooralahzadeh et al., 2021)	2021	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308		
PPKED (Liu et al., 2021b)	2021	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-		
Align Transformer (You et al., 2021)	2021	0.378	0.235	0.156	0.112	0.158	0.283	-	-	-		
NGYUEN ET AL (Nguyen et al., 2021b)	2021	0.495	0.360	0.278	0.224	0.222	0.390	-	-	-		
MDT+WCL (Yan et al., 2021a)	2021	0.373	-	-	0.107	0.144	0.274	0.384	0.274	0.294		
\mathcal{M}^2 Trans (CE) (Miura et al., 2020)	2021	-	-	-	0.111	-	-	0.463	0.732	0.567		
\mathcal{M}^2 Trans (EN) (Miura et al., 2020)	2021	-	-	-	0.114	-	-	0.503	0.651	0.567		
KGAE UNSUPERVISED (Liu et al., 2021d)	2021	0.221	0.144	0.096	0.062	0.097	0.208	0.214	0.158	0.156		
KGAE SEMI-SUPERVISED (Liu et al., 2021d)	2021	0.352	0.219	0.149	0.108	0.147	0.290	0.360	0.302	0.307		
KGAE SUPERVISED (Liu et al., 2021d)	2021	0.369	0.231	0.156	0.118	0.153	0.295	0.389	0.362	0.355		

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 102802.



- CNNs still very competitive w.r.t. Transformers in visual tasks
- ViTs report results slightly better or similar to CNNs, they are not outperforming them as in NLP
- Good results by combining both approaches

•

•

•

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical Image Analysis, 102802.



Modelos de Lenguaje

¿Qué es un modelo de lenguaje?

 Un modelo de lenguaje es un modelo que <u>asigna</u> <u>probabilidades a</u> <u>secuencias de palabras</u>




¿Qué es un modelo de lenguaje?

 La mayoría de los LMs actuales estiman la probabilidad condicional de la "siguiente palabra" dado un <u>contexto</u> previo

$$P_{\boldsymbol{\theta}}(\cdot | La \ silueta) \longrightarrow \begin{bmatrix} 0.001 \\ \cdots \\ 0.15 \\ 0.1 \end{bmatrix} \text{ felina}$$

$$cardiaca$$

$$cardiaca$$



Modelos de Lenguaje Neuronales

 Casi todos los LMs actuales son modelos de lenguaje neuronales, es decir, una red neuronal se usa para estimar los parámetros la distribución de probabilidad condicional



Attention is all you need: The Transformer

- Unlike computer vision where there is still no large performance gap between CNNs and ViT...
- In NLP, the Transformer has largely outperformed RNN/LSTM models



Attention is all you need: The Transformer

Self-attention





Online draft of "Speech and Language Processing" (3rd ed.) by Jurafsky and Martin https://web.stanford.edu/~jurafsky/slp3/

Self-attention versus RNN architectures

- The inherently sequential nature of recurrent networks makes it hard to do computation in parallel.
- These considerations led to the development of transformers, an approach to sequence processing that eliminates recurrent connections





• Encoder-only, Decoder-only, Encoder-Decoder





e.g. BERT

Encoder Example: BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp

2 - Supervised training on a specific task with a labeled dataset.



https://jalammar.github.io/illustrated-bert/

E-D Example: T5, FLAN-T5



Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "Text-to-Text Transfer Transformer".

#parameters: 60M (T5-small) – 220M



Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485-5551.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Text-only tasks in Radiology

- A NLP task in radiology is generating the <u>Impression</u> from <u>Findings</u> as a <u>summarization task</u>
- In BioNLP 2023 (ACL) there was a challenge about this task



Manual tags: Calcified Granuloma/lung/upper lobe/right Automatic tags: Calcified granuloma Comparison: Chest radiographs XXXX. Indication: XXXX-year-old male, chest pain. Findings: The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality. Impression: No acute cardiopulmonary process.

CT Abd/pelv	CT Chest	CT Head
15,989	12,786	31,402
CT Spine	MR Head	CT Neck
5,517	7,313	1,140
CT Sinus	Mr Spine	MR Abdomen
1,267	2,821	1,061
MR Neck	MR Pelvis	
230	253	

Table 1: Number of reports (findings/impression pairs) for each new modality/anatomy in the MIMIC-III summarization dataset, totaling 79,779 samples.



https://aclanthology.org/2023.bionlp-1.45/

Using LLM for impression summarization

- The team shs-nlp (Siemens) leveraged the *BloomZ* LM to create *RadBloomz (RBz)*
- BloomZ (300M to 176B parameters) is based on BLOOM (176B), BigScience Large Open-science Open-access Multilingual Language Model
- BLOOM has a Megatron GPT2 architecture



https://huggingface.co/bigscience/bloomz



BigScience Large Open-science Open-access Multilingual Language Model Version 1.3 / 6 July 2022

https://huggingface.co/bigscience/bloom

Karn, S. K., Ghosh, R., & Farri, O. (2023). shs-nlp at RadSum23: Domain-Adaptive Pre-training of Instruction-tuned LLMs for Radiology Report Impression Generation. *arXiv preprint arXiv:2306.03264*.

Using LLM for impression summarization

 An interesting result of the shs-nlp team is that <u>zero-</u> <u>shot RBz</u> worked better than after additional finetuning

Team	hidden testset	BLEU4	ROUGE-L	BertScore	F1-cheXbert	F1-RadGraph
shs-nlp		18.36	35.32	57.26	N/A	36.94
utsa-nlp		16.05	34.41	57.08	N/A	36.31
aimi	MIMIC-III	16.61	33.43	55.54	N/A	35.12
sinai		17.38	32.32	55.04	N/A	33.96
knowlab		13.23	32.02	55.64	N/A	33.39

Models	open test-set	BLEU4	ROUGE-L	BertScore	F1-cheXbert	F1-RadGraph
RBz-0shot	MIMIC III	17.33	33.93	55.49	N/A	34.93
RBz-ft		16.49	35.25	57.29	N/A	31.12
RBz-0shot	MIMIC CVD	25.32	47.48	63.61	74.34	49.00
RBz-ft	WIIWIIC-CAR	16.16	26.16	52.22	53.1	31.07

Karn, S. K., Ghosh, R., & Farri, O. (2023). shs-nlp at RadSum23: Domain-Adaptive Pre-training of Instruction-tuned LLMs for Radiology Report Impression Generation. *arXiv preprint arXiv:2306.03264*.

Lessons from BioNLP 2023 Report Summarization Challenge

General Trends

 Pretrained language models and vision-language models were commonly utilized, particularly BART, T5, OFA, BLOOMZ, BioBart, and SciFive

- Teams performed domain-specific fine-tuning using the provided
 training sets as well as external large-scale datasets
- Additional strategies that were successfully employed include:
 - · Freezing layers during fine-tuning to prevent overfitting
 - Model ensembling from multiple pretrained models and/or beam search sizes
 - Data augmentation with back-translation
 - Prompt tuning
 - Injection of linguistic information via a variety of pretraining tasks

 Modelos de Lenguaje usados: BART, T5, OFA, BLOOMZ, BioBart, SciFive

Additional strategies:

- Freezing layers during fine-tuning to prevent overfitting
- Model ensembling
- Data augmentation with back-translation
- Prompt tuning
- Injecting linguistic information via different pretraining trasks

Using a LLM as a surrogate expert labeler

- Another NLP tasks are sentence ranking and Natural Language Inference (NLI)
- But this requires too many annotations from experts
- Why not using LLMs as surrogate experts labelers ?



Figure 2: Fact extraction procedure for radiology reports.



Modelos Multimodales Visión-Lenguaje



Resuming the report generation task

¿How to use text and visual data to improve either or both representations?

P(granuloma within right lung |



???

Multimodal Deep Learning

multimodal adjective

health Millennium Institute for Intelligent Healthcare Engineering

mul·ti·mod·al (məl-tē-'mō-d^əl a) - tī-

: having or involving several modes, modalities, or maxima

multimodal distributions

multimodal therapy

Purpose: learn features over multiple modalities



Figure 1. Overview of the VATT architecture and the self-supervised, multimodal learning strategy. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the noise contrastive estimation to train the model.

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems, 34, 24206-24221.



If we consider two modalities, text and images:

- Retrieval (image <> text)
- Captioning (image -> text)
- Generation (text -> image)
- Visual question answering (image+text -> text)
- Multimodal classification (image+text -> label)
- Better understanding/generation (image+text -> label/text)



WSABI (Weston et al 2010), DeVise (Frome et al 2013), Cross-Modal Transfer (Socher et al 2013)



First models attempt to generate visual-semantic embeddings

Show and tell (Vinyals et al., 2015)

Show, attend and tell (Xu et al., 2016)



Image captioning was the first MM task using end-to-end DL for multimodality

https://web.stanford.edu/class/cs224n/slides/Multimodal-Deep-Learning-CS224n-Kiela.pdf



Visual BERTs: Supervised Multimodal Bitransformers



MMBT Kiela et al. 2019

Visual BERTs: VisualBERT, ViLBERT, LXMERT, etc.

CLIP (Radford et al. 2021)

Exact same contrastive loss as earlier, but.. Transformers and *web data*!





So many models

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EffcientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

Du et al. 2022 VLP Survey

https://web.stanford.edu/class/cs224n/slides/Multimodal-Deep-Learning-CS224n-Kiela.pdf

MedNeurIPS 2022: Two-stage conditional report generation

Two-stage Conditional Chest X-ray Radiology Report Generation

Pablo Messina ^{1,5,6}, José Cañete ^{2,6}, Denis Parra ^{1,5,6}, Álvaro Soto ^{1,6}, Cecilia Besa ^{3,5}, and Jocelyn Dunstan ^{4,5}

MedNeurIPS 2022: Two-stage conditional report generation



MedNeurIPS 2022: Two-stage conditional report generation

			NLP		Med. CheXpert (Macro)		(acro)	CheXpert (Mic		licro)		
ID	Model	B	R-L	C-D	Comp.	F1	P	R	F1	P	R	
	Other works											
1	Liu et al. [17]	0.192	0.306	1.046	-	0.180	0.313	0.126	0.334	0.634	0.227	
2	Chen et al. 2020 [7]	0.205	0.277	-	-	0.276	0.333	0.273	-	-	-	
3	Chen et al. 2021 [6]	0.208	0.283	-	-	0.303	0.352	0.298	-	-	-	
4	Lovelace et al. [18]	0.257	0.318	0.316	-	0.228	0.333	0.217	0.441	0.475	0.361	
5	Miura et al. [20]	-	-	0.509	-	0.304	0.361	0.360	0.563	0.499	0.646	
6	Nguyen et al. [22]	0.339	0.390	-	-	0.412	0.432	0.418	0.576	0.567	0.585	
7	Pino et al. [24]	0.094	0.185	0.238	-	0.428	0.381	0.531	-	-	-	
8	Kong et al. [14]	0.243	0.286	-	-	-	-	-	0.519	0.482	0.563	
			Our wor	k								
9	CRG(DN+TF) _{chexpert topics} : M	0.146	0.196	0.041	0.087	0.464	0.377	0.713	0.557	0.428	0.797	
10	CRG(DN+TF) _{chexpert topics} : M+I	0.146	0.196	0.041	0.087	0.469	0.388	0.678	0.569	0.448	0.781	
11	CRG(DN+TF) _{chexpert topics} : M+I+Ch	0.146	0.196	0.041	0.088	0.463	0.384	0.689	0.568	0.446	0.783	
12	CRG(DN+TF) _{chexpert topics} : M+I+Ch+C14	0.146	0.196	0.040	0.088	0.463	0.386	0.702	0.564	0.440	0.785	
13	CRG(DN+TF) _{chexpert topics} : M+I+Ch+C14+V _{test}	0.145	0.195	0.040	0.088	0.467	0.386	0.712	0.569	0.439	0.811	
14	CRG(DN+TF) _{chexpert topics} : M+I+Ch+C14+V _{all}	0.145	0.195	0.041	0.088	0.462	0.383	0.700	0.571	0.444	0.800	
15	$CRG(DN+TF)^{medtok, fve, ft}_{chexpert topics}$: M+I	0.146	0.197	0.040	0.086	0.477	0.392	0.693	0.575	0.449	0.799	
16	CRG(ViT _{CLIP} +TF) ^{medtok, fve, ft} _{chexpert topics} : M+I	0.150	0.199	0.040	0.087	0.472	0.389	0.653	0.582	0.464	0.779	
17	CRG(DN+TF) ^{medtok, vmf, ft} mined topics predicted by ensemble: M+I	0.102	0.184	0.031	0.116	0.448	0.400	0.568	0.588	0.487	0.743	

MedNeurIPS 2022: Is the NLG considering the input image?

Table 2: Visual encoder results on the test split of MIMIC-CXR. For CRG models, CheXpert metrics for the Transformer decoder when conditioned on CheXpert topics are included. Cohen's Kappa measures the agreement between visual encoder and Transformer.

		Mined	Topics	CheXpert (visual encoder)				CheXpert (transformer)		
ID	Model	F1 (macro)	F1 (micro)	ROC- AUC (macro)	ROC- AUC (micro)	F1 (macro)	F1 (micro)	F1 (macro)	F1 (micro)	Cohen's Kappa
1	CRG(DN+TF): M	0.213	0.413	0.758	0.823	0.473	0.578	0.465	0.556	0.703
2	CRG(DN+TF): M+I	0.208	0.409	0.750	0.821	0.465	0.579	0.476	0.570	0.767
3	CRG(DN+TF): M+I+Ch	0.206	0.392	0.763	0.821	0.472	0.576	0.474	0.569	0.785
4	CRG(DN+TF): M+I+Ch+C14	0.209	0.401	0.765	0.823	0.478	0.582	0.470	0.563	0.806
5	CRG(DN+TF): M+I+Ch+C14+V _{test}	0.212	0.405	0.761	0.826	0.483	0.587	0.476	0.570	0.785
6	CRG(DN+TF): M+I+Ch+C14+V _{all}	0.210	0.405	0.762	0.829	0.481	0.587	0.475	0.572	0.806
7	CRG(DN+TF) ^{medtok} : M+I+Ch+C14+V _{test}	0.216	0.422	0.765	0.829	0.486	0.593	0.473	0.569	0.781
8	CRG(ViT _{CLIP} +TF) ^{medtok, fve, ft} : M+I	0.219	0.396	0.743	0.823	0.471	0.590	0.474	0.582	0.811
9	TC(DN+ChEmb+Bilstm) ^{e=191} : M+I+Ch+C14+V _{all}	0.233	0.443	0.715	0.804	0.450	0.566	-	-	-
10	$TC(ChEmb+Bilstm)^{e=74,ft}$: M+I+Ch	0.230	0.497	0.744	0.804	0.463	0.562	-	-	-
11	TC(DN+ChEmb+Bilstm) ^{e=180,ft} : M+I+Ch	0.234	0.516	0.740	0.813	0.465	0.573	-	-	-
12	$TC(DN+ChEmb+Bilstm)^{e=568,ft}: M+I+Ch+C14+V_{all}$	0.239	0.507	0.741	0.817	0.466	0.579	-	-	-
13	TC Ensemble	0.310	0.603	-	-	-	-	-	-	

BioNLP 2023 Report Summarization challenge: OFA/CheXOFA

The winner of the report summarization challenge in BioNLP 2023 used OFA and finetuned it



OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework



https://arxiv.org/pdf/2202.03052.pdf

ihealth Millennium Institute for Intelligent Healthcare Engineering Flamingo -> Flamingo-Med



Med-Flamingo 2



This is the code repo for the Med-Flamingo paper.

https://github.com/snap-stanford/med-flamingo

https://web.stanford.edu/class/cs224n/slides/Multi modal-Deep-Learning-CS224n-Kiela.pdf Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., ... & Leskovec, J. Med-flamingo: A multimodal medical few-shot learner. July 2023. *arXiv preprint arXiv:2307.15189*.



LLaVA: Large Language and Vision Assistant



https://llava-vl.github.io/

LLaVA-Med: Large Language and Vision Assistant for BioMedicine @

Visual instruction tuning towards building large language and vision models with GPT-4 level capabilities in the biomedicine space.

• [June 1] (A) We released LLaVA-Med: Large Language and Vision Assistant for Biomedicine, a step towards building biomedical domain large language and vision models with GPT-4 level capabilities. Checkout the paper



LLaVA-Med was initialized with the general-domain LLaVA and then continuously trained in a curriculum learning fashion (first biomedical concept alignment then full-blown instruction-tuning). We evaluated LLaVA-Med on standard visual conversation and question answering tasks.

https://github.com/microsoft/LLaVA-Med

Google Palm -> MedPalm 2

Towards Generalist Biomedical AI

Tao Tu^{*, ‡, 1}, Shekoofeh Azizi^{*, ‡, 2},

Danny Driess², Mike Schaekermann¹, Mohamed Amin¹, Pi-Chuan Chang¹, Andrew Carroll¹, Chuck Lau¹, Ryutaro Tanno², Ira Ktena², Basil Mustafa², Aakanksha Chowdhery², Yun Liu¹, Simon Kornblith², David Fleet², Philip Mansfield¹, Sushant Prakash¹, Renee Wong¹, Sunny Virmani¹, Christopher Semturs¹, S Sara Mahdavi², Bradley Green¹, Ewa Dominowska¹, Blaise Aguera y Arcas¹, Joelle Barral², Dale Webster¹, Greg S. Corrado¹, Yossi Matias¹, Karan Singhal¹, Pete Florence², Alan Karthikesalingam^{†, ‡,1} and Vivek Natarajan^{†, ‡,1}

¹Google Research, ²Google DeepMind



Modality	Dataset	Metric	SOTA	PaLM-E (84B)	Med-PaLM M (Best)
Chest X-ray	MIMIC-CXR	Micro-F1-14 Macro-F1-14 Micro-F1-5 Macro-F1-5 F1-RadGraph BLEU-1 BLEU-4 ROUGE-L CIDEr-D	44.20% [65] 30.70% [65] 56.70% [66] N/A 24.40% [14] 39.48% [65] 13.30% [66] 29.60% [67] 49.50% [68]	$15.40\% \\ 10.11\% \\ 5.51\% \\ 4.85\% \\ 11.66\% \\ 19.86\% \\ 4.60\% \\ 16.53\% \\ 3.50\%$	53.56% 39.83% 57.88% 51.60% 26.71% 32.31% 11.50% 27.49% 26.17%

[cs.CL] 26 Jul 2023



- GPT4-V is capable of generating radiological reports
- OpenAl does not currently recommend its use

Prompt	GPT-4V Response 1	
Which side of the brain is the hydrocephalus located on in this image?	The hydrocephalus in this image appears to be located on the left side of the brain.	Wrong Answer
	GPT-4V Response 2 The hydrocephalus in the image appears to be located on the right side of the brain. GPT-4V Response 3	Correct Answer
	The hydrocephalus in this image appears to be located on the right side of the brain (from the viewer's perspective). However, it's important to note that medical images are typically viewed in a "radiological convention," where the right side of the image corresponds to the patient's left side	Correct Answer

Figure 7: Examples of GPT-4V's unreliable performance for medical purposes.

... we do not consider the current version of GPT-4V to be fit for performing any medical function or substituting professional medical advice, diagnosis, or treatment, or judgment...

Modelos Generativos: RoentGen

- Se han hecho populares modelos como DALLE-2 y Guided Difussion para generar imágenes a partir de texto
- RoentGen es un sistema que genera imágenes de Rayos X de pecho dado el texto de impresión



https://stanfordmimi.github.io/RoentGen/



Discussion

Need for more public data

- An important challenge is the need to generate public clinical databases in order to accelerate advances.
- This need is even greater in languages other than English

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 31 Mar, 2022

We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world - see also our <u>NeurIPS2022 paper</u>

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev

Dataset	Year	Image Type	# images	# reports	# patients	Used by papers					
Report datasets											
IU X-ray [27]	2015	Chest X-Ray	7,470	3,955	3,955	[15, 35, 39, 47, 60,					
						65, 66, 83, 84, 87,					
						90, 114, 117, 126,					
						136, 142, 143, 145-					
						148, 154]					
MIMIC-CXR [67, 68]	2019	Chest X-Ray	377,110	227,827	227,827	[90]					
PadChest ^(sp) [18]	2019	Chest X-Ray	160,868	109,931	67,625	None ⁽⁵⁾					

Biases and Discrimination

- Using generative models like GPT or RoentGen can be useful for data augmentation
- Risk: increasing biases and not creating improvements in populations different from those of the original data



Figure 1: Experimental setup and ensemble architecture. Fairness approaches are indicated in dotted boxes.

Sensitive	Train	Test	~ ~ ~
Group	Count	Count	% of Test
F	7940	1415	44.0 %
М	9708	1778	56.0 %
ASIAN	408	60	1.9 %
BLACK	1658	285	8.9 %
HISPANIC	521	107	3.3 %
OTHER	2655	459	14.4 %
WHITE	12406	2282	71.5 %
Government	356	74	2.3 %
Medicaid	1362	205	6.4 %
Medicare	9857	1757	55.0 %
Private	4946	932	29.2 %
Self Pay	133	33	1.0 %
UNKNOWN	994	192	6.1 %

ihealth Millennium Institute for Intelligent Healthcare Engineering Learning with imbalanced Data



Fig. 1: Long-tailed distribution of thorax disease labels for the proposed NIH-CXR-LT (left) and MIMIC-CXR-LT (right) training datasets. Values by each bar represent log-frequency, while values in parentheses represent raw frequency. Textured bars represent newly added disease labels, which help create naturally long-tailed distributions without the need for artificial subsampling.

Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study

https://gholste.me/publication/holste-dali-2022/holstedali-2022.pdf

Table 1: Long-tailed learning methods selected for benchmarking grouped by type of approach ("R" = Re-balancing, "A" = Augmentation, "O" = Other). "RW" = re-weighted with scikit-learn weights [22], "CB" = re-weighted with class-balanced weights [4].

Method	R	A	0	Method	R	A	0
Softmax (Baseline)				CB LDAM-DRW [2]	\checkmark		
CB Softmax	\checkmark			RW LDAM [2]	\checkmark		
RW Softmax	\checkmark			RW LDAM-DRW [2]	\checkmark		
Focal Loss [16]	\checkmark			MixUp [29]		\checkmark	
CB Focal Loss [16]	\checkmark			Balanced-MixUp [6]	\checkmark	\checkmark	
RW Focal Loss [16]	\checkmark			Decoupling–cRT [12]	\checkmark		\checkmark
LDAM [2]	\checkmark			Decoupling $-\tau$ -norm [12]	~		\checkmark
CB LDAM [2]	\checkmark						

LLMs & Hallucinations

health Millennium Institute for Intelligent Healthcare Engineering

- Inability of AI models to recognize "when they don't know" (uncertainty)
- Very active research area, but problem is still unsolved

Prompt:

Write a radiology report for this image as an expert radiologist.



GPT-4V: Report: Date: 01/01/2021 Exam: X-ray of the right hand and wrist Technique: AP and lateral views were obtained Findings: The bones of the right hand and wrist appear inta

Findings: The bones of the right hand and wrist appear intact and in proper alignment. There is no evidence of fracture or dislocation. The joint spaces are preserved and there is no evidence of degenerative changes. No soft tissue abnormalities are seen. Impression: Normal X-ray of the right hand and wrist.

Evaluation from Professionals: Missed the obvious distal radial fracture Correct identification of the type of study

Is the problem of report generation solved?

- Not yet
- Even if it is solved there are other challenges for supporting medical practice:
 - Follow-up of patients over subsequent sessions
 - Diagnosis -> Prognosis, Precision Medicine
 - Recommend need for other image modalities or exams


Challenges

- Expand to other pathologies & images' types (MRI, CT-Scan, Echotomography, etc.)
- Generate to other languages beyond English
- Deal with multimodal input (text, images, videos, tabular, Genomic data, etc.)
- Interpretable AI: explain predictions of results
- Improve generalization (OOD samples): transfer learning, metalearning, etc.
- Evaluation: NLG metrics vs actual clinical diagnostic



* Questions to denis.parra@uc.cl

