AUTOMATED MEDICAL IMAGE REPORT GENERATION

Gasimova, A.*, Montana, G.†, Rueckert, D.*
*Biomedical Image Analysis Group, Department of Computing, Imperial College London
† Imaging and Biomedical Engineering Clinical Academic Group, Biomedical Engineering Department, King's College London

INTRODUCTION

Gathering manually annotated images for the purpose of training a predictive model is far more challenging in the medical domain than for natural images as it requires the expertise of qualified radiologists. We therefore propose to take advantage of past radiological exams and formulate a framework capable of learning the correspondence between the images and reports, and hence be capable of generating diagnostic reports for a given X-ray examination consisting of an arbitrary number of image views. We demonstrate how aggregating the image features of individual exams and using them as conditional inputs when training a language generation model results in auto-generated exam reports that correlate well with radiologist-generated reports.

LEARNING FROM MEDICAL REPORTS

Learning to Read Chest X-Rays [1]:

- Applied Neural Image Caption Model of Vinyals et. al [2] to chest X-rays and **MeSH term annotations** of reports created by radiologists.
- MeSH term annotations are difficult to automate, require a trained language model.

TandemNet [3]:

- Joint attention model over image regions and text, trained on bladder cancer histopathology images and reports.
- Pathologists asked to write reports according to a **template**, learning framework is therefore limited.

Novelty of this Project:

- Use of **past** radiological examinations and corresponding **raw, free-text reports.**
- Framework handles **arbitrary number of input images.**

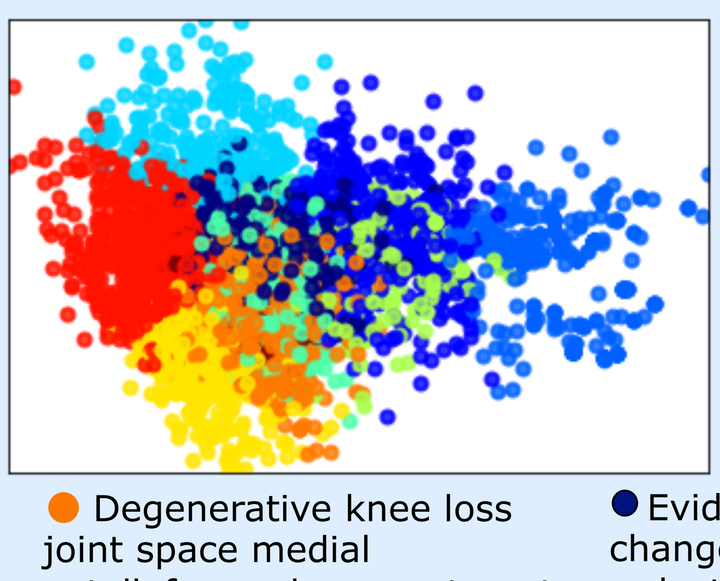
DATA

The knee X-ray dataset has been extracted from the PAC system of St Thomas Hospital (part of Guys and St Thomas NHS Foundation Trust) and has been fully anonymised to remove sensitive patient information.

- 330 knee X-ray exams collected over 2 years.
- Each exam consists of a textual report and one or more X-ray images (left/right knee, taken from different views: anteroposterior (AP), lateral (L) and skyline (S)).
- The reports vary in length between 2 and 145 words, avg. 30, and between 1 and 16 sentences, avg. 2.7 per report.
- The X-ray images vary in sizes between 420 × 650 × 3 and 3056 × 3056 × 3.

Report Embedding Clusters

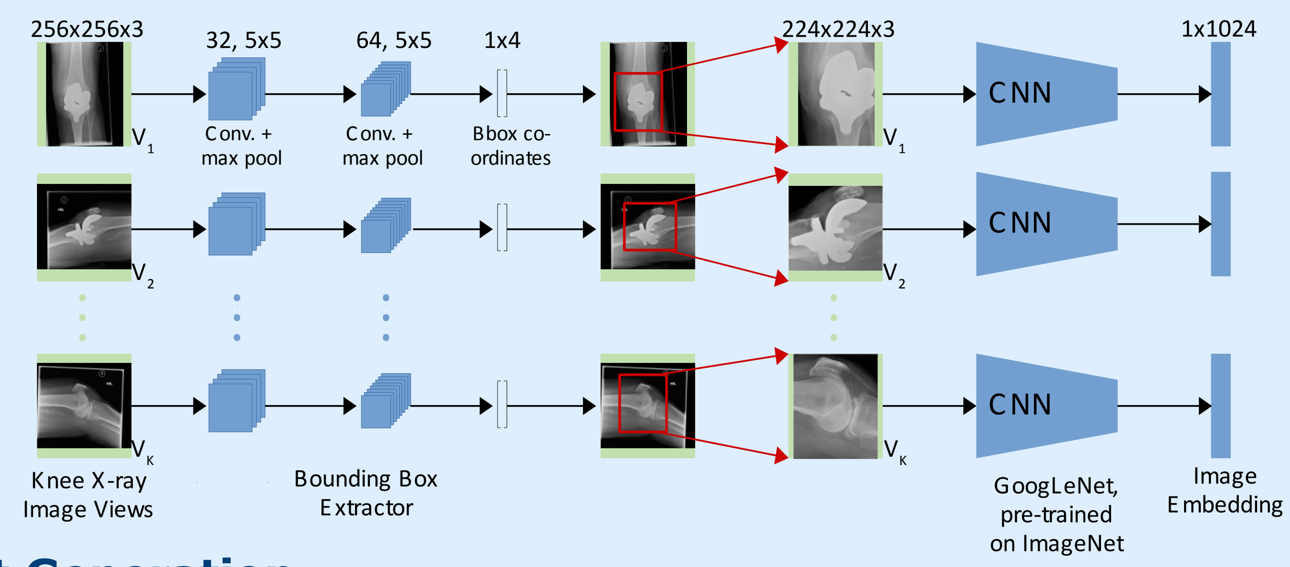
Generated using InferSent [4] language model. Pathologies in reports generally fall in the categories of *degenerative change, joint space narrowing, fracture, prosthetic loosening*, and *normal*. Common modifiers include severity *mild/moderate/significant* and locations *medial/later/ patellofemoral compartments.*



FRAMEWORK

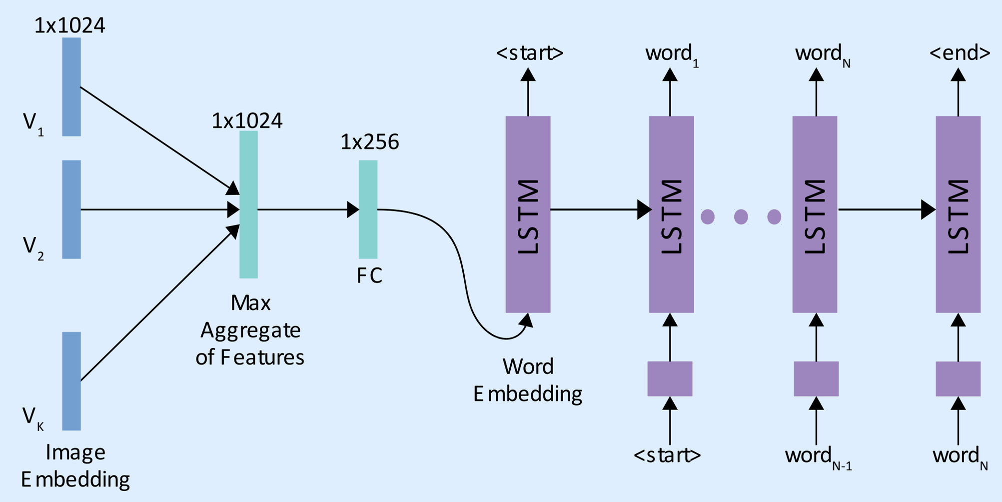
Image Modeling

A CNN BBox regressor built and trained on a subset of the training images (231) to detect the knee joint. Image features are extracted from the last spatial average pooling layer of GoogLeNet, pre-trained on ImageNet.



Report Generation

Max-aggregated images features for each exam are input at time step t=0, words in the report input at consequent time steps.



Training

Report Generation Model trained by minimising the negative log-likelihood:

$$L(S,I) = -\sum_{t=0}^{N} \log\left[p(P_t = T_t | \mathrm{CNN}(I), P_0 \ldots P_{t-1})\right]$$

where p is the probability that the predicted word P_t equals the true word T_t at time step t given aggregated image features *CNN(I)* and previous words $P_0 \ldots P_{t-1}$, and N is the LSTM sequence length.

Dataset was augmented eight-fold by:

- Random cropping the images from 256 × 256 to 224 × 224
- flipping the images along the vertical axis
- shuffling the sentences in the reports.

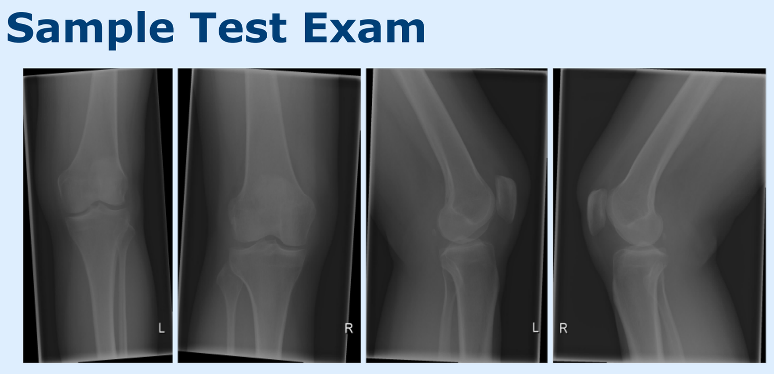
RESULTS

	BLEU -1		BLEU -2		BLEU -3		BLEU -4		METEOR	
	tr	te	tr	te	tr	te	tr	te	tr	te
Baseline, single image input	42.2	33.2	13.3	5.7	3.8	1.9	1.3	1.1	26.7	22.2
Max-aggr. of image features	60.7	**40.4**	32.6	**10.1**	19.4	**2.6**	12.3	**1.2**	41.1	**35.7**
Max-aggr.+Bboxes	38.9	37.4	11.3	7.1	3.4	1.1	1.2	0.2	28.3	28.9

BLEU-n/METEOR Metrics

Modified form of n-gram precision commonly used for evaluating image captioning and machine translation.

Sample Test Exam



True Report: Joint spaces articular surfaces appear preserved. Significant degenerative erosive change seen.
'Good' Prediction (B1=87.5): Joint spaces articular surfaces appear preserved bilaterally.
'Poor' Prediction (B1=28.6): Joint space narrowing medial compartments bilaterally.

Conclusion

Preliminary results look promising as the auto-generated reports correlate well with true reports, and we hope to train the model on additional knee X-ray exams as these become available to us. Further developments to the model can be made by incorporating the knowledge of the view-type of each image, keeping them as separate inputs, and finding correspondence between image regions and parts of text.

References
[1] Shin, H.-c., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016). Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. pages 2-5.
[2] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156-3164.
[3] Zhang, Z., Chen, P., Sapkota, M., & Yang, L. (2017, September). TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 320-328). Springer, Cham.
[4] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A., (2017) Supervised Learning of Universal Sentence Representations from Natural Language Inference Data arXiv preprint arXiv:1705.02364