1. **Image captioning model using attention and object features to mimic human image understanding:**
   Authors: Muhammad Abdelhadie Al-Malla, Assef Jafar and Nada Ghneim
   Dataset used: MSCOCO & Flickr30k
   Models used: Encoder (Xception, YOLOv4), Decoder (GRU)

   Paper shows comparison of baseline model(without object features) and proposed model(with object features).

   YOLOv4 is used to extract object features and added them in Xception model. After adding the Image features, the CIDEr score increased by 15.04%.

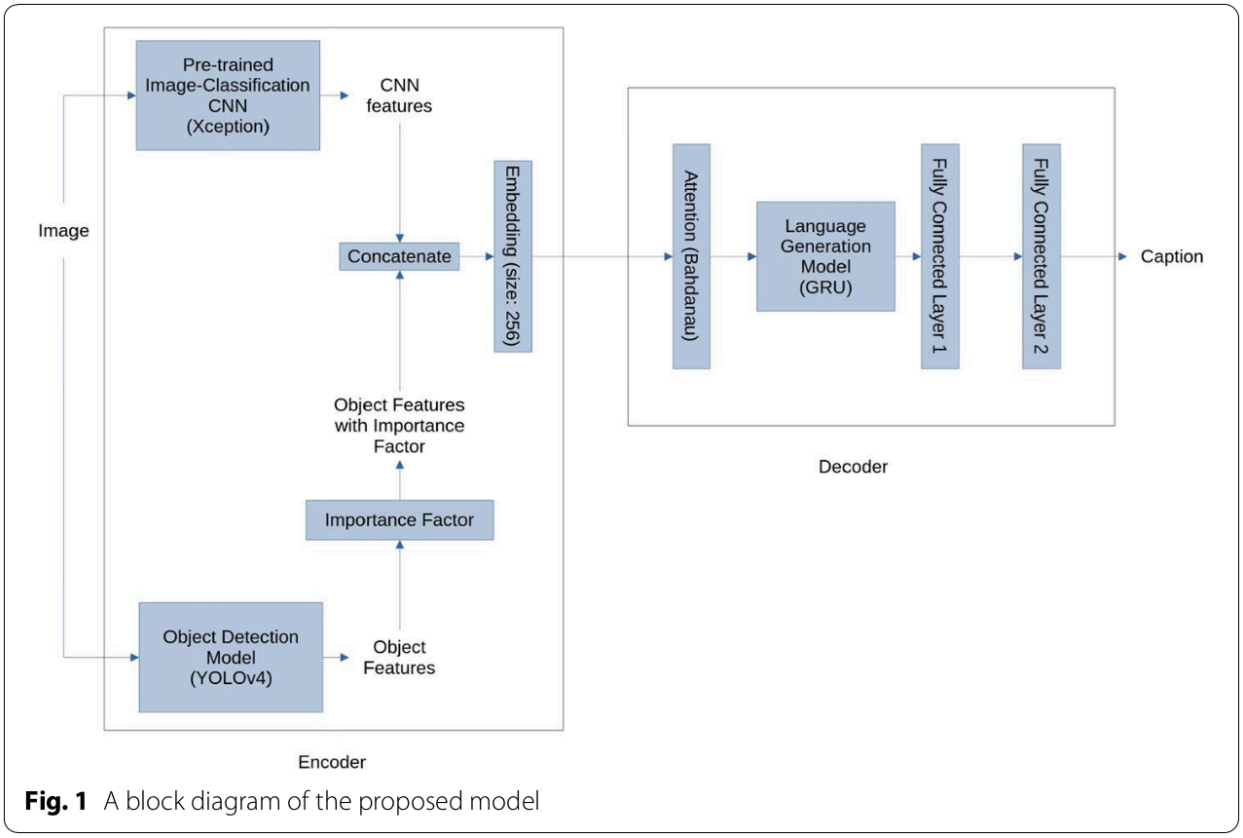

Fig. 1 A block diagram of the proposed model

   (Flickr30k, 20 epochs, loss function-Sparse Categorical Cross Entropy)

2. **Adaptive Attention-based High-level Semantic Introduction for Image Caption:**
   Authors: XIAOXIAO LIU and QINGYANG XU
   Dataset used: MSCOCO an Flickr30k
   Models used: CNN (ResNet), LSTM
   Batch size-10, training iterations- 1,50,000, training time- 16 hours

Table 2. Comparison Among Models on Flickr30k

Model	B-1	B-2	B-3	B-4	METEOR	CIDEr	ROUGE-L
NIC [11]	66.3	42.3	27.7	18.3	–	–	–
(SS+RA)-Ensemble [31]	64.9	46.2	32.4	22.4	19.4	47.2	45.1
Soft attention [8]	66.7	43.4	28.8	19.1	18.49	–	–
Hard attention [8]	66.9	43.9	29.6	19.9	18.46	–	–
SCN-LSTM Ensemble of 5 [50]	74.7	55.2	40.3	28.8	22.3	–	–
Att-reginCNN+LSTM [3]	73.0	55.0	40.0	28.0	–	–	–
Stimulus and concept driven [35]	66.3	43.7	29.2	21.1	–	–	–
VSDA [5]	68.1	49.8	35.7	25.6	20.8	53.2	47.4
SGA-BR [7]	69.0	50.9	36.8	26.4	21.5	–	–
Adaptive attention [28]	67.7	49.4	35.4	25.1	20.4	53.1	–
hLSTMat [38]	73.8	55.1	40.3	29.4	23.0	66.6	–
Ours	**86.4**	**68.3**	**51.5**	**38.2**	**26.9**	**83.4**	**57.5**

3. **An Overview of Image Caption Generation Methods:**
   Datasets considered: MSCOCO, Flickr8k Flickr30k, PASCAL
   Challenges discussed: Generating complete natural language sentences like a human being, Grammatical correctness, inconsistency.
   Feature extraction methods discussed:
   - Handcraft Features with Statistical Language Model
   - Deep Learning Features with Neural Network

   Second method shows good results(good sentence structure)

4. **Image Captioning: Transforming Objects into Words:**
   Authors: Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares
   Datasets: MSCOCO

Algorithm	CIDEr-D	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L
Att2all [21]	114	-	-	34.2	26.7	55.7
Up-Down [2]	120.1	21.4	79.8	36.3	27.7	56.9
Visual-policy[15]	126.3	21.6	–	38.6	28.3	58.5
GCN-LSTM [29][1]	127.6	22.0	80.5	38.2	28.5	58.3
SGAE [27]	127.8	22.1	**80.8**	38.4	28.4	**58.6**
Ours	**128.3**	**22.6**	80.5	**38.6**	**28.7**	58.4

   - BLEU-1 score of 80.5.
   - Used PyTorch, Tesla V100 GPU. Time taken for training- 1 day.
   - ResNet-101 is used as CNN for feature extraction, LSTM for caption generation.

- Input image–>ResNet-101–>LSTM–>Training with Self-Critical Reinforcement Learning–>caption(output)

5. **Image captioning with referent objects attributes:**
   Authors: Seokmok Park1 & Joonki Paik
   Datasets used: RefCOCO, COCO
   ResNet-101 fine tuned on VisualGenome
   Dataset splitted into 70(training), 20(testing), 10(validation)

   (i) Object detection: This module enables the system to search for relevant visual content based on user queries.
   (ii) Visual grounding: The visual grounding module aims to establish a connection between textual queries and specific objects or regions in the visual content.
   (iii) Scene graph generation: This module generates a structured representation of the relationships between objects in the scene, capturing their interactions and contextual information.
   (iv) Image captioning: The image captioning module generates descriptive captions that accurately convey the content and context of the visual input.

Models	BLEU-4	METEOR	ROUGE	CIDEr
SCST[24]	25.3	25.7	50.1	131.4
Up-Down[38]	25.5	26.8	53.2	137.1
ClipCap[13]	33.5	30.4	–	124.1
GRIT[39]	**38.2**	**30.3**	55.7	142.9
RefCap	33.2	29.7	**56.2**	**143.7**