Survey on News Image Captioning

Deepesh Tank , 18075017, CSE IIT BHU

3 December 2021

1 Paper 1 [1]

1.1 Importance

News image captioning is a hard task in the field of CV (computer vision). This paper deals with captioning that incorporate both textual and visual description of the article-image corpus. Earlier work in this field deals with the extraction of a certain set of keywords depicting the pair information but did not deal with the importance of text features. Named entities like Delhi, Max, Alex, Russia, etc are very hard to tell by seeing the mere image therefore the traditional conventional method of captioning fails to provide effective results.

No previous work is done in this type of captioning therefore this paper provides a useful path of exploration in integrating the text of an article with an image to produce a useful caption. These types of generated captions are very useful in assisting journalists to make headlines for news without reading the entire article by themselves and also give them a third-person perspective of the news. This type of perspective makes readers curious and therefore increases the overall production growth.

Another useful application of this paper can be decoding subtitles from one language to another sometimes merely translating the subtitles to another language doesn't perfectly depict the actual meaning due to the difference of slang, dialect, etc., therefore, captioning involves extracting information from moving frame images and current subtitles can be used to derive best subtitles of another language. Slight modification to the model taking image frame by frame along with subtitles can achieve this task.

This paper also consists of a comparison of different types of embedding and variants. This comparison study can be very useful in setting benchmarks for other transformer models. Finally, the model presented in the paper can be modified according to the need to perform other captioning tasks that take both image and text information into consideration.

1.2 Methodology

We have a pair of an article and images related to specific news. The objective is to produce a caption for the news that ideally constitutes the interpretation of the mentioned pair. Paper has carried through the task using the transformer model.

In starting, large articles are pruned to a fixed size length of at most 416 words. In addition to this non- Ascii characters are removed from the article. The character dot is used as a delimiter. The vocabulary of 32,000 words is generated using Byte-Pair-Encoding. Here, text input is represented as a sequence of tokens (p^1, p^2, p^3, p^m) therefore, the final output is expected to be returned in a format of tokens (q^1, q^2, q^3, q^m). Thus making the objective a sequence to sequence problem and therefore the transformer model is the best-suited architecture for the implementation.

The whole architecture is comprised of three main parts: picture encoder, article-picture encoder, and decoder.picture encoder takes the picture as an input and converts it into a vector x (feature vector $x \in R^s$).

$$x^{(image)} = \text{CNN}^{(image)}$$

CNN(comparison study on two CNN models ImageNet and Places 365) is used as an encoder and s is the dimension of the output vector of this encoder. The combined representation of the pair is done by the article-picture encoder. A matrix of dimension d $\times$ n is given as starting input. The column vectors of this input matrix combine aspects of positional encoding and token embedding. Two substiute components namely the encoder and picture-attending module are stacked to form N layers. The first module is mapping from R^{s*n} to R^{s*n} which is done by transformer model through its multi-head self-attention layer and subsequently processed by its feed-forward neural layer.

The other module is mapping from R^{s*n} to R^{s*n} which is done by transformer model through its multi-head target-source attention layer and subsequently processed by its feed-forward neural layer. In this way, the first module and second module are the same as the encoder-decoder part of transformer architecture respectively. The decoder is comprised of M layers, taking encoder output as key and values. These inputs are used for scaled dot-product attention.

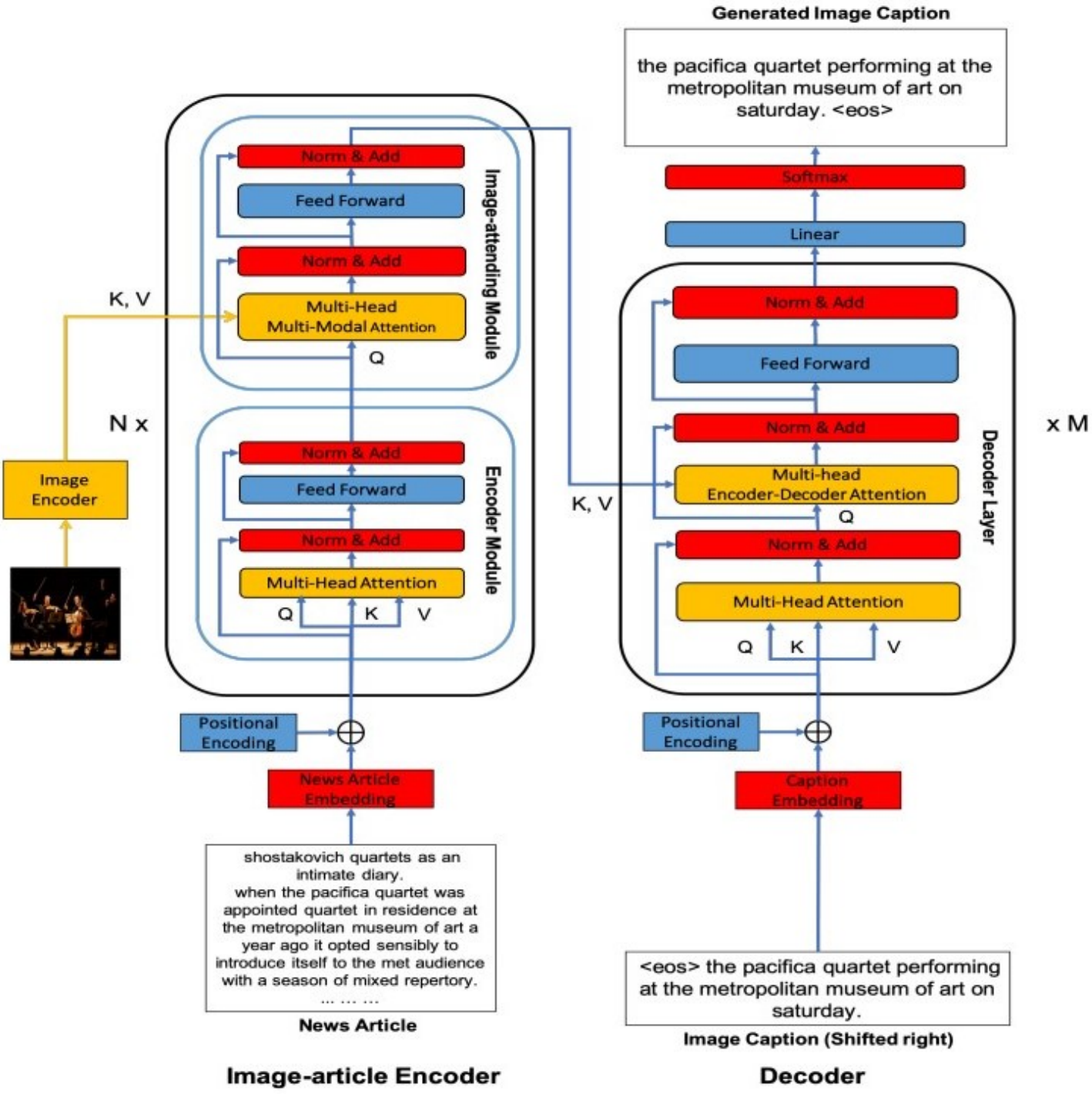

Figure 1: Model Architecture.

This encode-decoder-type architecture ensures that the overall model preserves useful textual and visual data. The attention mechanism ensures selectively taking only important data from both image-article pairs to generate the most accurate caption. Article embedding is done in multiple ways for comparison studies namely:

- taking the mean of Glove embedding for every sentence,
- taking the weighted mean of Glove embedding for every sentence,
- tough to beat baseline,
- named entity incorporation using attention mechanism,
- named entity incorporation based on context.

Drop-out layers of rate 0.3 are added after token embedding and positional encoding combination, after normalization, and before residual connections. Another dropout layer of 0.1 rate is applied for weight attentions.
Comparison study includes 6 variants of the main model namely

- text-only,
- image only imagiNet,
- image only Places 365,
- multimodal imageNet, multimodal places 365,
- multimodal combined imageNet-places 365.

Model is trained over GoodNews dataset tuning the above-mentioned settings. Experimental results reveal multimodal variants are best than others and have less notable differences among themselves. Hyperparameters are set as s = 512 , attention heads H = 8, N =3 and M = 6.. the cross-entropy loss is taken as the main loss function with label smoothing of 0.1 value. The model is tested on 5 evaluation metrics. News image captioning tasks deal with a lot of named entities therefore mention metrics may not capture the best of the model. Therefore, a new evaluation metric named CoverageNE came into the picture and is defined as the following.

$$\text{CoverageNE} = \mid \text{E}_{output} \cap \text{E}_{original} \mid / \mid \text{E}_{original} \mid$$

E represents a set of named entities. Industrial grade Natural language Processing tool SpaCy is used for collecting named entities from the text.

1.3 Drawbacks

- Not a perfect model results are heavily dependent on the choice of dataset and the choice of embedding and variants.
- very heavy model, a single run takes over days to run on a powerful GPU.
- For getting the best result over a specific dataset, we need to run the dataset over all possible combinations of settings, this will further increase the running time. Therefore not feasible for automating caption for day-to-day news.
- Different type of dataset requires a different type of pre-processing which is itself a very time-consuming task.
- Since the entity is incorporated into the embedding there may be grammatical errors in the output although it is containing all the useful information.
- Captions are something that is judged by human interpretation therefore there is a limit up to which the model can be optimized.
- Full image encoding also takes non-objective things into consideration like background color, hue, contrast therefore it affects the results very significantly.
- The unnecessary text associated with the article like repeating some sentences, associated notes, results in memory overhead.
- Since the architecture is based on encoder-decoder some important information is always lost midway in transitioning from high to low-level dimension vector.
- Transformer models are not able to context and content when the sentences of the articles are very long.

2 Paper 2 [2]

2.1 Importance

Retrieving data from a large chunk of multimodal sources of information is a very hard and time-consuming job. News occupies a major portion of this chunk providing useful data about important person's lives and their stories. This type of work concerned with a person requires automatic identification of a concerned person.

In the past decade significant groundbreaking work in done in this field of expertise. The majority of such work is only related to a single language article. When international news comes into the picture authors from all around the world write the articles in their native language. This poses challenges to the traditional focus personage in news images. This paper aims to solve such challenges by introducing a novel method that can be applied to such news articles containing more than one language. Such type of method can lay the foundation for new globalization stone in the field of international media.

The paper is tested over a large data set of Google and Baidu images. The results clearly show that the model overperforms another modal quite significantly in terms of error percentage and human interpretation. Traditional methods require translating the news into a single language and then performing the operation over it. This results in error accumulation from the mentioned two-step process as the translation doesn't always give perfect results. The elimination of the first step improves the accuracy of focus personage. The conversion from a two-step process to a single-step process removes additional time overhead and the use of AP clustering further adds to time benefits. The paper opens new doors in the field of multi-lingual feature extraction and the underlying architecture's scalability ensures tweaking to obtain the best results.

2.2 Methodology

This paper proposed a novel method that can automate the process of focus person identification in new containing multiple languages. The method used here also take the fact into consideration that name shared by the image in the attached captions are having to be associated with the face and in addition to this, the count of the matching face images related to the mentioned name is considerably greater than the count of another face images. This type of conjecture is acceptable for multi-phonetics news images.
In the first step to focus personage, method explore for new-image caption set related to her/his name. After this, the paper applies a face diagnostic algorithm to images of news to find all the possible face images. The mentioned process is suitable for multi-phonetics news images. For effective creation of training inputs, for each and every possible face image, the method divides the face images into two types:

- Contain images with a single face and caption also have a single name.
- Rest of the remaining images

The first type of category is operated to train positive samples by the algorithm of affinity propagation Clustering and the positive sample is operate in training RCNN. Concurrently the second type of category is operated in RCNN for testing and classifying purposes.

Caption preprocessing
For multi-lingual, a name can be written in a different manner. For example, Deepesh, Shēnyuān(chinese) are all related to a single person. Therefore In this paper, we develop a vocabulary by translate and collecting the set of dissimilar names wield for the single person and succeeding it , taking the convergence to identify face image related with unlike names of a matching person.

Face detection
First ASMs algorithms are performed over the collection of news images. This results in the location of the feature points of the face that subsequently result in the feature point location of facial entities like mouth, nose, tongue, etc. For aligning the face in a particular direction eyes are used as a static feature point. Wrapping is done in three color channels RGB and finally, these channels are combined to get complete wrapped faces.

Getting facial description is a hard task because of several imaging factors like light, shadows, posture, facial expressions, blurriness, wrinkles, etc. the method obtains LGBPHS of the facial images and then perform KPCA to lower down the dimension of the obtained optical attributes.

AP Clustering
The use of AP clustering is due to its lower error proportion and less running time than other available methods. The method performs AP clustering over the first type of face images. It considers the likeliness of information points as input and then concurrently uses all information points as inherent epitomes. For likeliness, Euclidean distance is used as a parameter which is defined as below.

$$S_{ab} = \text{squareRoot}(\sum_{k=1}^{K} (I_{ak} - I_{bk})^2) \quad a,b \in 0,1,2,...,N$$

S_{ab} denotes the likeliness of I_a / I_b Images where N is the count of the first type of images. K represents the dimension of the feature. For the first type, a face image with maximum likelihood is related to focus personage. AP clustering is utilized to obtain maximal clusters that ensure the deletion of blurry and noisy face images. For the multi-phonetics photos of focus

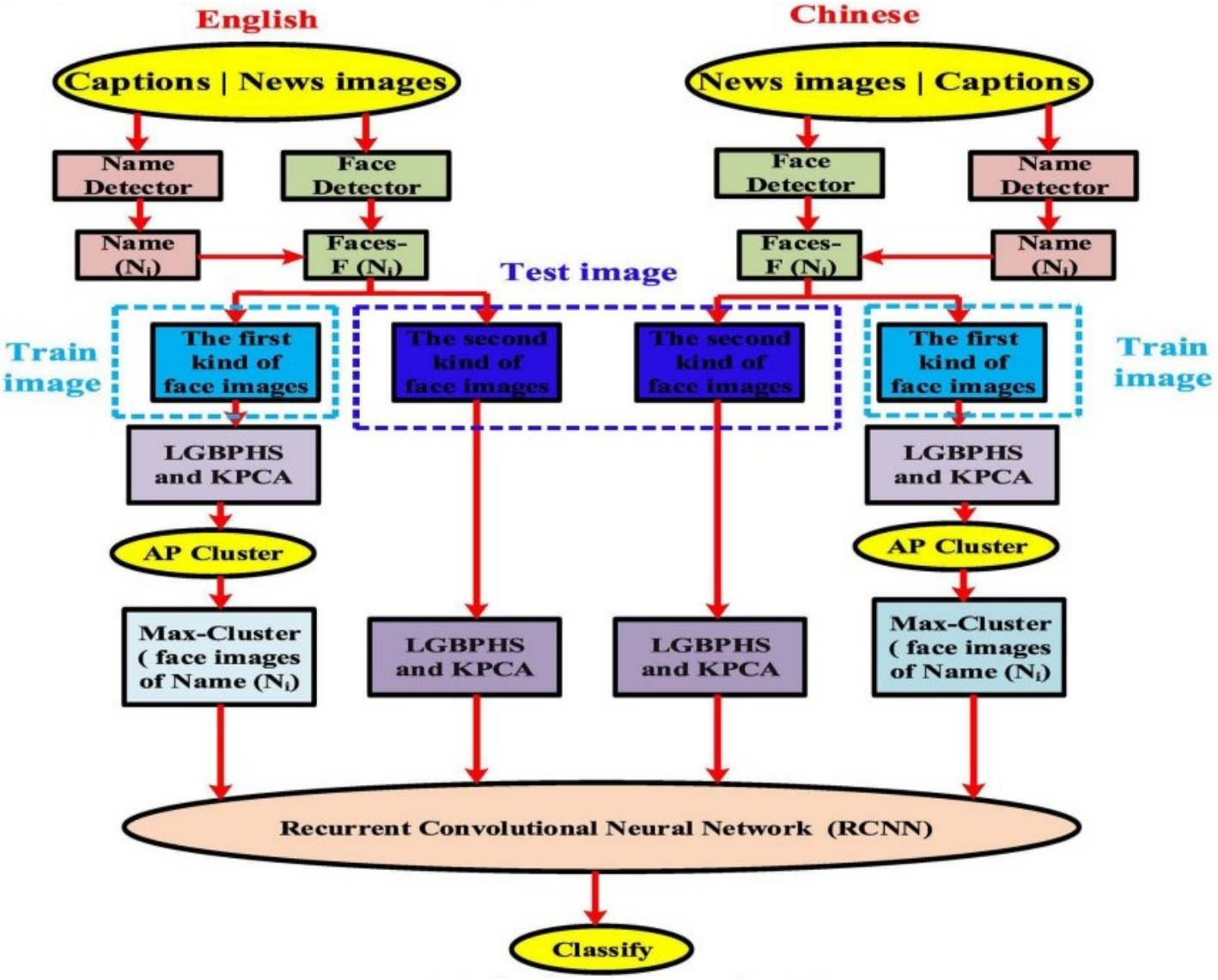

Figure 2: The flowchart of the model.

personage, affinity propagation is utilized single-lingual photos individually. This results in more than one maximal cluster of obtained focus personage. Ensembling all these clusters results in the terminal cluster and all the face images in the terminal cluster are made use of as positive samples for feeding the RCNN model.

RCNN

In the paper, RCC is comprised of one Conv Layer, 3 RCLs, 3 maximum pooling layers, and a single softmax layer. Conv Layer is a feed forward conv layer not having any recurrent attachments. There is only a feed-forward attachment among adjacent RCLs and the last RCL's output goes through a global maximum pooling layer taking the largest among all the output maps. The resulting output is considered as a vectorial portrayal of the image. The last layer is categorizing the features into C kinds as follows.

$$p_k = \exp(\ {w_k}^T q\ ) \ /\ (\sum_{k'} \exp({w_{k'}}^T q)) \quad k = 1,2,3,...,C$$

p_k represents the predicted likelihood related to the k^{th} kind, and q is the vectorial feature obtained just before the softmax layer.

2.3 Drawbacks

- The proposed model doesn't give the best results on distorted, noisy, and blurry images.
- The model doesn't have a single pathway of execution instead it divides the face images into two categories and applies different methods over them. A new type of use-case may result in further categorization.
- AP clustering is less time-consuming when compared to similar methods but in a general sense, it has a higher time complexity($O(n^2 \times \log(n))$) therefore not suitable for large datasets.
- Small changes to the models can result in drastic changes as the AP algorithm is very sensitive to the choice of input features.
- The use of RCNN in the model leads to problem of not performing well on long sentences of articles.
- Error rate and Time consumed in training the model heavily depend on the language used. Therefore for some languages, the model may not generate the best results.
- As the number of languages increases in the article, the consumed time also increases drastically therefore there is a limit up to which the model is feasible for use.
- The method may not work on cropped face images as it uses eyes to wrap the image in a desired conical shape. In addition, it also has challenges in identifying the person who has covered his/her eyes in the new image.
- The model also suffers from an inherent lingual problem when two different names may be written as same in another language.
- Time consumed(1.5 hr) in training is almost double from Su's method(0.6 hr) although the error rate is smaller than the other.

3 Paper 3 [3]

3.1 Importance

Occurrence of named entities in an article overall uplifts the level of difficulty in image-text matching tasks. Over the past decades, numerous amounts of progress is made in this field but most of them are not directly able to predict named entities in the caption. The main reason is the deficiency of inter-relation analysis between named entities. Therefore, this paper gains importance as it exploits the relationship between them to predict a meaningful caption for news articles.

The model is compared with a variety of baseline models. The model overperforms all the base models overall evaluation matrices. Therefore the model can be used in industrial applications for example in NLP tools, Artificial Intelligence. The model can also be used as a standard one to compare other models training to accomplish the similar task of image-text matching. The model used is best in analyzing the semantic differences between the images and articles. Since the model primarily fills the placeholders with the appropriately named entities that are most likely in terms of probability. Therefore slight modification in the model can also be used to fill the missing information or gaps in text articles. This type of work identifying missing words in the text is a huge challenge for forensic scientists. Therefore our model can be helpful in such situations.

The model is a general-purpose model slight variation in the structure can lead to finding a new semantic pattern that is hard to discover using the human brain. Our model finds its application in search engines where the variation of the model can prove search recommendations based on the incomplete text imputed by the user. The field of news-image matching is very large day by day progress is made all around the world. Various papers are published daily but only a few of them have a wide range scope of utility and the method used in this paper covers a significant portion of this image-text matching field.

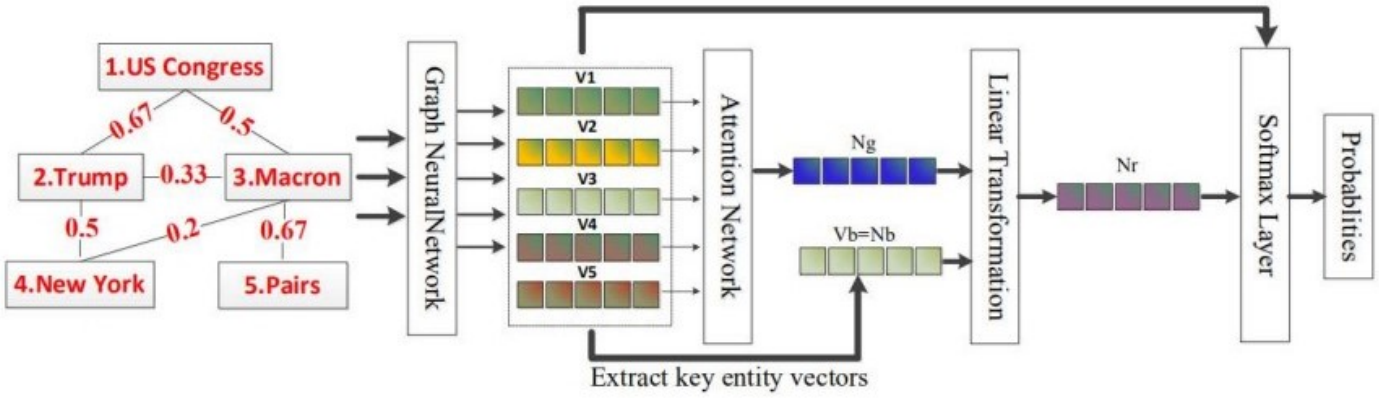

Figure 3: The NKD-GNN architecture.

3.2 Methodology

The pair focus to identify the regularity of news-image and text by producing captions containing named entities. The model proposed by the paper is divided into four stages. The First Stage of the model is to build up a news knowledge graph. This graph establishes a relation between entities present in the article. The second stage produces a draft containing the news objects and placeholders. The third stage is the main proposed NKD-GNN method that examines the entire correspondence and deduces the nondirect connection among the named entities in the build-up graph. The method then chooses the best option for every placeholder in the produced draft. In the end, Hybrid Co-Attention Network (HCAN) comes into the picture to identify the uniformity of the caption with article text and named entities in it.

Producing News Draft caption

The paper produces the draft caption which has placeholders for named entities in addition to their tags. This stage is based on the state-of-the-art model. Encoder Decoder-based architecture is utilized with CNN and LSTM playing the role of encoder and decoder respectively. CNN here contains 19 Conv layers trained on the imageNet dataset and finally, CNN uses its last fcc layer as the main encoding unit. LSTM is utilized to decode the encoding to act for captions. This way of decoding generally doesn't have many named entities and has a difference in semantics. Therefore to solve the problem, the draft is produced utilizing WordNet. There is a limit of four types of producing placeholders i.e. organization, venue, building, and person. For each tagged placeholder we select the best fitting option to fit in the respective category.

Creating News Knowledge Graph

The main focus of the proposed model is to establish and identify the entire correspondence between named entities in the article of the news. The general method is to exploit the knowledge graphs. To create such a graph, the named entities from the article must be known and past this, the weight of correspondence between entities must be established. The industrial-grade NLP tool SpaCy is used for this purpose. The dataset used is TopNews.the model construct set $P = p_1, p_2, \ldots, p_m$ of all named entities extracted from the dataset. The model constructs another set $W = w_1, w_2, \ldots, w_m$ to represent the entities that are in the same sentence or in some way appear next to each other. The associated weight is calculated by the formula.

$$J_e = f_{phpt} \;/\; \max(f_{ph}, f_{pt})$$

In this formula $e \in E$, J_e defines the edge e's value. e is the edge connecting p_h (f_h) and p_t(f_t) named entities in the graph. f_{phpt} is the synchronized repetition of the mentioned entities. The more closed the entities appear repeatedly the larger the weight their edge will have and vice versa. In the graph, a particular node resembles the entity and the edge resembles weighted correspondence.

Graph NN

To study the entire correspondence and deduce the nondirect uniformity among entities (named ones) in the graph, the NKD-GNN method is proposed with the following architecture. The architecture is comprised of 4 stages. In the first stage, each point on the graph has to sum all edges and point data in the graph. This way the entity vector is generated. The attention-based mechanism is used to give weightage to the edges of the graph thus calculating the global representation vector N_g of the graph. In the graph, the key entity has the highest number of edges. To calculate the representation of graph N_r vector, the linear transformation over the combination of N_b entity and N_g global vector is taken. Finally, to calculate the probability for every single entity the multiplication of its vector v_i with N_r is done.

Prediction

The paper calculates the z score for every single entity after getting the graph representation Nr vector by taking the product of vector vi and Nr (graph representation vector).

$$Z_i = {N_r}^T V_i$$

We take the softmax output of Z to get model vector Y.

$$y\wedge = \text{softmax}(Z_i)$$

Here, the z represents the score of node, y∧ depicts the probabilities of entities that can occupy the placeholder of the draft caption. After this, the entity having the highest value of probability score is used to fill the placeholder. As a consequence of this step, given the draft caption ' <PERSON> in a black coat and hat sitting in a <PLACE>', we get ' Deepesh in black coat and hat sitting in a hotel.' For the placeholder that is not able to get any entity we fill them with the general name associated with the placeholder. For example, using the word place to fill placeholder <PLACE>. The model uses cross-entropy as the objective function to minimize the loss.

$$\text{LossCalulation}(y\wedge) = -\sum_{i=1}^{m} y_i \log(y_i \wedge) + (1\text{-}y_i)\log(1\text{-}y\wedge)$$

Here, y vector represents the one-hot encoding of the most suitable entity in the graph.

Computational Technique for consistency
The paper defaults the news-image along with textual consistency when the template including the named entities meets a sole sentence of the text article. Here, the paper follows the HCAN model as the task is a text-matching one.
The HCAN model is comprised of three core parts:

- Hybrid encoder exploring three categories of encoders namely wide, contextual and deep.
- Relevance consistency component having external weights that are used for analyzing term consistency signals.
- Semantic consistency component having co-attention functions for context-mindful representation learning.

As it is well known that HCAN also accounts for relevant matching of semantic of two sentences. In this way, it defeats the layout dissimilarity of the sentences between caption including named entities and text article.

3.3 Drawbacks

- The absence of named entities for some of the placeholders decreases the overall quality of the new image caption.
- General text is placed in a placeholder in absence of entity over a large number of such missing blanks will lead to an increase in general text in placeholders. For example person, place, etc will occur more frequently that will deteriorate the overall quality of the new image caption.
- Model is tested over GoodNews dataset which contains a large percentage of the one-word captions that will impact the overall results very badly.
- Dimension the new graph depends on the number of the present entity in the article and it can affect the model's performance for odd datasets
- Accuracy range from 78.4 to 85.7 percentage over different dataset it shows that the model is not generalized for all datasets.
- The model can't unsheathe the analogical(metaphorical) news image therefore its uses depend on the genre of news articles. for example 'five-star red flag' points to China etc.
- Model's performance decrease over a large number of entities in the sentence or for long sentences as the attention mechanism is not effective for longer text.
- Some named entities can have different meanings and different types for example India can be a place(<PLACE>) as well as women(<PERSON>) therefore the model fails to work in such situations. The model needs to categorize the same word twice which will lead to a significant change in the model structure which can then lead to bad results on normal entities.
- The use of an attention mechanism leads to an increase in the number of trainable parameters as a consequence time taken for training is very large.
- Not useful for conceptual abstract articles and news as the number of entities in such articles is very low.

4 Paper 4 [4]

4.1 Importance

Over the last decade, a huge amount of progress is made in the job of automatic image captioning. However, the current captioning tools basically specify the objects in the image. therefore, the generated draft caption doesn't contain the real-world knowledge of the named entities and the interrelation among them for example some information about a famous person, place, and communities present in the image. In opposite, humans decode the image in a very certain way by providing some valuable information about the named entities of the image. Therefore, the task of generating a human-like caption is gaining importance.

In this paper, the focus is on captioning in such a way that it incorporates some real-life data about the image or the named entities present in the image with the help of the available contextual article associated with the image. The paper proposed a novel method that generates human-like captions of the news image by taking advantage of the relevance of semantic properties of the news named entities. First of all, named units are extracted from the semantic relations between the visual and textual features and then a sentence analysis algorithm comes into action to selectively extract the textual features from the text article then the paper uses an entity algorithm that is based on a knowledge graph to identify the overall relations among different sets of named entities. The model primarily fills the <placeholders> with the appropriately named units that are most likely in terms of probability. Therefore slight modification in the model can also be used to fill the missing information or gaps in text articles. This type of work identifying missing words in the text is a huge challenge for forensic scientists. Therefore our model can be helpful in such situations.

The paper is tested over the GoodNews datasets and Breaking news. The paper is compared with several baseline models that include hard-attention, Adaptive, Group cap, etc. The 5 standard evaluation metrics are used. The results show the proposed model outperforms the existing baseline models on all evaluation parameters making the paper successful in the task of news-captioning. In order to have a fair evaluation, human manual evaluation is done and the results show that our model is getting good results overall. The outperforming results on the 5 standard evaluation metrics show that our model can be used as a baseline model for comparing new models that are being proposed for the news-image captioning task.

4.2 Methodology

The main task of the paper is to generate human-like captions(descriptions) for the photos by taking background information into consideration. The background information is provided in the form of news articles. The overall method consists of four main steps. In the first step, the main goal is to get sentence-based representation so the text is preprocessed in starting. This step is performed at the starting point because the textual data similarity is easy to make at basic granularity.

In the second step, the paper defines an algorithm based on sentence co-relation as the textual data similarity. This algorithm is used for matching the news images and articles in a joint space including global semantics association. In the third step, the model is trained on the select image-article corpus. The named entities in the sentences are taken over by the placeholders in the stage of preprocessing. In this way, we get a descriptive caption with all the words and placeholders replacing named entities. The draft caption producer follows the encoder-decoder architecture with a CNN model as an encoder and LSTM as a decoder.

In the fourth and the final step, in the order to replace the placeholders with accurately named entities, the paper used an entity linking algorithm, which is basically a news knowledge graph algorithm. Then the certain named entities are linked which is removed from the semantic related article's sentence to the outside knowledge bases. Then the best suited named entities for every placeholder are chosen.

Preprocessing
Preprocessing is done in three steps:

- short sentences are useless for the model so we remove sentences less than five words.
- To handle the longers sentences the paper uses a compression method(Stanford dependency parser). It preserves the essential descriptive part of the sentence like subjects, verbs, and objects.
- Identification of the named entities in the sentences is itself another NLP task therefore we use SpaCy to identify them. After this, the paper links the identified entities to the DBpedia datasets to give reference entities. This ensures we have a large set of named entities to replace the placeholders in the final template.

Sentence level presentation
To ensure the preservation of the textual information, we reduce the granularity level of representation from article to sentence level. Similarly, the dimension matrix of the article news expressions at the sentence level is less than the word/phrase level. To encode the words of the sentences, the Continuous bags-of-words model (CBOW) embedding is used. For encoding the sentence we use the average aggregation algorithm.

$$\mathrm{Si} = 1/((\mathrm{N}_i) \sum_{j=1}^{Ni}) \quad \mathrm{w}_j \ , \mathrm{i} = 1, 2, 3, \ldots.., \mathrm{M}$$

In the above formula, Si denotes the ith sentence in the article consisting of the Ni words. The CBOW embedding is represented as wj where j is from 1 to Nj. Each word has an impact on the outcome therefore we take a weighted mean of vectors that are based on the property of the SIF(smoothed inverse frequency). The final improvised algo reduces the influence of non-useful words.

$$\mathrm{S}_i = 1/(\mathrm{N}_i) \sum_{j=1}^{Ni} \mathrm{a}/(\mathrm{a} + \mathrm{p}(\mathrm{wj})) * \mathrm{w}_j$$

The weight of the word is $\mathrm{a}/(\mathrm{a} + \mathrm{p}(\mathrm{w}_j))$ and $\mathrm{p}(\mathrm{w}_j)$ is the frequency. After this, the frequency-inverse document frequency (TF-IDF) is applied acting as weightage associated with each sentence to complete the whole article encoding process.

Finding contextual data of image entities
The news article is consists of several sentences but only a quite handful of them are actually contextually associated with the news image. For relating the sentences to the respective image for the semantic association, the paper applies a contextual data similarity algorithm that is specifically based on the deep canonical interrelation analysis.

First of all, representation XI is computed for the image by the VGG19 Convolution layer model. The output of the last fcc layer is a 1000-dimensional sparse vector. Then, sentence representation is computed. A news article is represented by $Y_A \in R^{DxM}$, D = D_w = 1000 and M is the total count of the sentences. Matrix X_I of the concerned image is zero-padded to the domain of Y_A and M - 1 columns of zeros are added to it. X_I and Y_A serve as the input points for the sentence analysis algorithm.

Generating template caption
The template generating task is similar to the image captioning job in which the training set is comprised of the news articles and images associated with the article. Only the news-related text is selected from the article. Type indicating placeholders are taken place of named entities following state of art captioning model. Vgg19 model trained on ImageNet is used for image encoding and the last fcc layer serves as an output on encoding and for decoding we use LSTM. Image vectors are set as hidden states initially. At the timestamp t, the model architecture generates the probabilities of words/placeholders on the basis of resultant output at t - 1 state and the hidden state.

4.3 Drawbacks

- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- The use of LSTM increases the training time and subsequently slows down the overall task.
- The absence of named entities for some of the placeholders decreases the overall quality of the new image caption.
- Not suitable for complex news where the content text is not directly related to the image.
- This model can't be used in certain fields as it is not able to give good results in the unorthodox news genre like satirical news, philosophical articles, etc.
- Not useful for conceptual abstract articles and news as the number of entities in such articles is very low.
- A weight-based mechanism is needed to be present in assigning weight to each image and news representation as some articles may be more inclined towards text meaning rather than image representation.
- Model is tested over GoodNews dataset which contains a large percentage of the one-word captions that will impact the overall results very badly.
- The model overperforms other baseline models but stills very far behind manual evaluations.
- The value of evaluation metrics differs heavily over different types of datasets therefore model may give poor results over some datasets.

5 Paper 5 [5]

5.1 Importance

In the recent years, there is a general interest in researching the similarity between image and text on the basis of their shared semantics. There are a large number of breakthroughs in tasks dealing with captioning and image retrieval. Most of these breakthroughs learnings are largely dependent on the big train dataset of photos that are associated with human annotations which gives some related or background information related to the concerned image. This paper deals with such tasks and raises it up to the next level to discover more complex cases in which there are fewer relations between the image and text part of the training datasets. The main focus is on captioning the news article type dataset where textual content has an ambiguous relationship with the associated image. The paper introduces an adaptive CNN-based model which shares a range of jobs namely article illustrations, source detection, geolocation information of the article, etc.

The proposed algorithm for the task of article illustration is Deep Canonical Correlation in addition a new objective loss function is proposed for the geolocation. This function is based on the great circle distance. Along with this, A new novel dataset called BreakingNews is proposed which contains around 100k articles of news. These articles contain images, captions, text, meta-data(geolocation information, author information, etc) that are used to explore the complex relationships between themselves.

The results show that the proposed dataset is well suited for various deep learning exploration tasks and can be used as a baseline comparison for different deep learning architectures. Besides handling the image captioning job, the BreakingNews dataset is intended to solve new rising challenges like media agency detection, estimating GPS coordinates. If anyone wants to take their combined vision and language development to the next level, this dataset is an excellent place to start. The resemblance between photographs and text in BreakingNews is not as straightforward as it is in other datasets, which means

that the objects, attributes, and activities of the photos may not pop up obviously as words in the text content. The overall results are very good but still, there is a large scope of improvement. The dataset can be further modified or even extended with some extra annotations to solve some other deep learning problems like visual question answering/ text summarization etc.

5.2 Methodology

The paper proposes a novel dataset of articles including images, comments, location data, and captions that are used to judge CNN and LSTM frameworks on a set of different jobs. The method is based on the state of art methodology of deep learning, therefore the dataset is a base stone to research the present limitation of these methods, the dataset is very useful when working with images containing text.

Preprocessing the dataset
The paper adds images from google with articles using title as query input to Google images. Manual verification proves that this approach is more precise than the use of full text as a search query. Among the search result, the top five images with reasonable size are selected and are linked to the article. After this, the images are annotated with the comments and number of shares by users. We also take author, date of publication, likes, dislikes, and other attributes into account.

Text Representation
Here, the paper uses the bad of words and Word2Vec embeddings. BoW is the most popular text representation format and is shown its abilities in different publications. It requires a set of vocabulary, that is entrenched as distinctive lemmatized tokens derived from the training information which display greater than L times in the articles. This will lead to D_b dimension BoW vectors. The j_{th} indexed dimension of this vector is presented by t_j log $(M/c_j + 1)$ where t_j represents the frequency of the j^{th} token, c_j refers to the document frequency and M is the count of the training articles. The most common way is to shorten the vector on the basis of the inverse document frequency. It is been proved that performance generally improves monotonically with the count of the retained dimensions.

Word2Vec embeds every word in a concrete-valued vector that ensures that the semantics properties are preserved. It uses a two-layer solution in which words/phrases are embedded based on their meaning and defined as a range that spans from past to future words. 2 methods are there to learn the representations: cbow model stands for 'Continuous bag of words' in which the objective is generating a word when the meaning or context is given and the second one is the 'Continuos skip-gram model' in which the objective is vice-versa of the former. In the negative sampling function, The target word has to be differentiated from the state of random negative words, this objective function is used as a reliable alternative to the softmax objective function. The paper compare both the objective function and found out that the skip-gram method is better in use for the paper.

Firstly, the model is trained, and this model is used to embed each article by combining the expression for every and each word and symbol in a matrix format with D_w dimensions times the count of tokens in articles. Here, D_w is the length of the space of embeddings.
From this proposed matrix, the paper research on three representations:

- Full matrix as input to CNN layers.
- Mean OR
- Max with respect to the sentence dimension to construct a unique Dw vector for every single article associated with the news image.

Another alternative would be to first apply pooling for each and every sentence independently and subsequently for all the sentences of the article in a single document vector.

Image representation
The image representation used in the paper is based on pre-trained deep convolutional neural networks. The paper follows the state-of-art methodology developed for the purpose of object recognition that includes 19 CNN layers(VGG19). The paper calculates the activation of the last relu layer of the VGG19 CNN and places recognition CNN. Every activation consists of 4096-dimensional sparse vectors and is proved to perform significantly better in various vision jobs. The paper also l_2 normalizes the two vectors and adds them to make a third image expression that is of the size 8192.

CNN architecture
The article is denoted by D_w x n_i matrix . D is the dimensionality of the space of embedding and n_i is the total count of the token available in the i^{th} article. This matrix is further zero-padded to the count of tokens presented in the longest article. The activations of the CNN layers are max pooled with each channel and the resultant 256-dimensional vector is then modified into 64 channels in an fcc layer. This is done before non-linearity is applied to the model. The study shows that the relu function is numerically more useful in terms of gradient calculations. Dropout layers are used to prevent overfitting in the overall model architecture. The optimal value of the dropout ratio is found to be 0.1 for our model. The gradient of the loss function is used to compute and backpropagate to update the CNN model parameters.

Caption generation

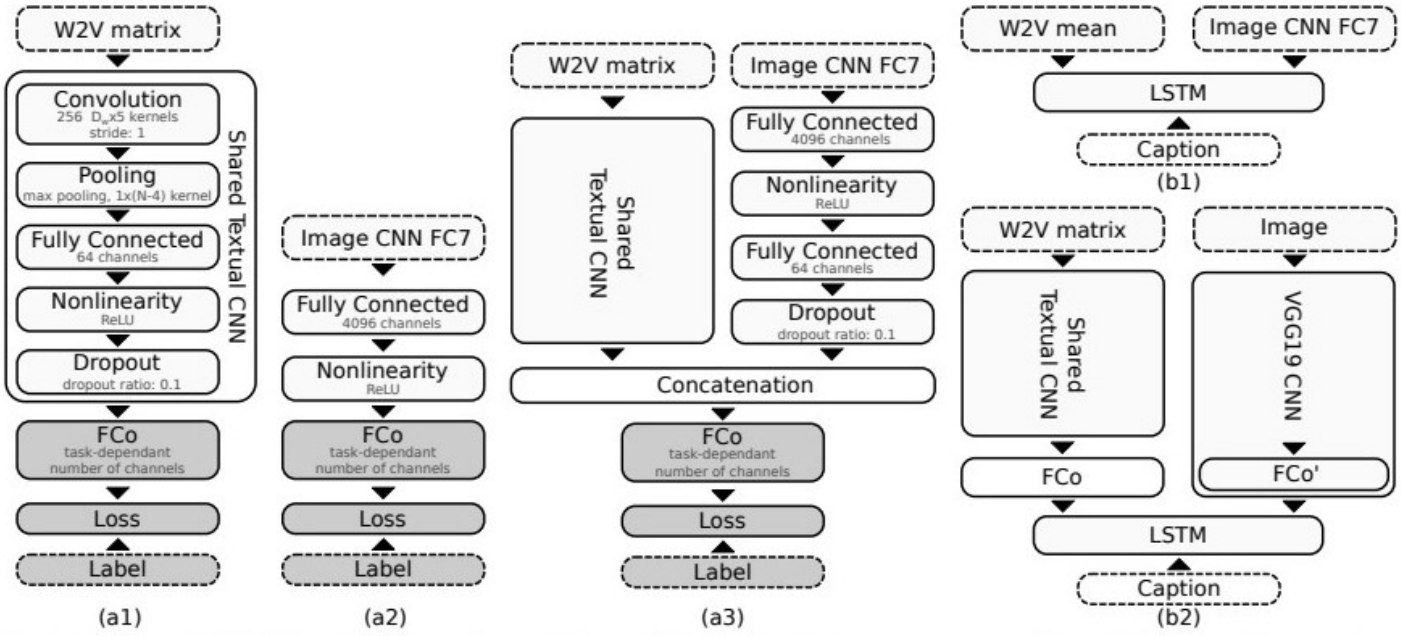

Figure 4: CNN and LSTM architecture for caption generation.

In this study, we take the following approach of the caption generation that is based on LSTM architecture. LSTM with permanent features: The job of news captioning is different from the traditional one in two forms. First, the caption of news is less tightly linked to what is described in the image that making this task very challenging. To solve this challenge, approaches are made to enhance the image descriptions, and actions similar to the attention mechanism are not used as a baseline. Second, news captions take both image and article into consideration rather than image alone. To solve this issue, the paper takes the mean of embedding matrix as permanent article representation and VGG19/Place activation as image representations. The LSTM is trained using one of the representations or the addition of both image and article representations.

5.3 Drawbacks

- The use of LSTM in the model architecture increases the training time of the model. This results in additional time overhead.
- The fixed value of the dropout ratio leads to bad results(captions) in certain news image pairs.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- Taking the additional information from the Google images is not a better way to solve the issue of incomplete information for some articles.
- The number of likes and shares extracted from the social networks are heavily manipulated therefore it may result in bad captions for the news image.
- The top 5 captions from the Google images may be very similar or exact duplicates of each other, therefore, it may result in overfitting of the model.
- Word2vec model cannot able to handle unknown words and the news articles contain a significant set of such words therefore word2vec model is not practically feasible for news image captioning tasks.
- Incorporating different properties like the number of likes, locations increases the complexity exponentially therefore there is a limit up to which the model can work properly.
- Named entities are the most important part of news image captioning and this paper doesn't deal with them, therefore, this paper is not good for news image captioning tasks.
- Since the architecture is somewhat similar to encoder-decoder some important information is always lost midway in transitioning from high to low-level dimension vector.

6 Paper 6[6]

6.1 Importance

Over the past decade, the amount of digital data present on the internet has increased exponentially. Browsing images in large-scale hetero datasets is an important challenge that has generated large attention in recent years. Over a large number

of search engines over the internet retrieve the image without taking their content into account. Therefore most of the search engines can't able to find relevant images associated with the queried text. This paper aims to solve such problems by generating textual descriptions for the images. This paper is related to the challenge of solving automatic generating headlines for news images. Other fields of applications include image and video extraction, build up of suitable machines that help a visually disabled person to access visual information.

The proposed model of the paper takes advantage of the large resource of images present on the internet along with the fact that a lot of them are captioned and information associated with them are available to access. The model is able to generate the captions from the dataset of images and articles of the news. The proposed novel method has its application in a variety of fields namely computer vision, natural language processing, video retrieval, image retrieval, development of machines for news media management, building tools for the visually disabled, lightning-fast breaking news delivery, etc. The key feature of the model is to allow both textual and pictorial modalities to affect the generation task. This is ensured by the image annotation technique that adds information in terms of keywords. The results prove that pictorial data is necessary for content selection. Picking the most effective sentence from the article as a caption is not a good method. The result shows that the generated captions are grammatically correct and manage to retain the essence of the image along with the article text. The captions generated are similar to those written by the journalists.

The paper primarily focuses on caption generation exploration. In addition to this, the model architecture can also be applied in other kinds of information namely image sharing sites, life-science publications that traditionally have graphical illustrations along with detailed textual information. There is also a large scope of improvement in the model and can be used for further in-depth exploration of similar complex tasks. For example, an infinite number of topics can be allowed to develop a nonparametric version that learns the number of optimal topics, the model can use spatial relationships among different parts of the image instead of just local features for representing the images, the model provides flexibility for experiment with other features like a particular section of the text article, exploiting syntactic data more directly(used in a phrase-based model taking attachment likelihood in consideration), improving grammar by dependency graph, etc.

6.2 Methodology

The main problem statement for the paper is: given a news photograph I and its related text article D, generate a descriptive caption C that perfectly depicts the meaning or essence of the photograph I. Therefore the training set consists of pair of photographs, article, and caption. Testing uses the former two to predict the later one.

Data validation
The BBC dataset is used for training purposes. The dataset mainly focuses on two purposes. To begin, the model utilizes the photos, captions, and related articles as training data to develop an photo annotation architecture that will obtain explanatory keywords for the photographs. The caption generating model will then be guided by these keywords. Second, the captions generated by humans will be taken as a golden standard for both the picture annotation architecture and the end-to-end caption predicting mechanism. In the first situation, stopwords will be removed or eliminated and the caption will be treated as a bundle of content words.

Modeling
The model architecture is comprised of dual stages. First is the selection of content that investigates what the photographs and related articles are about. The second is surface realization which decides how to verbalize the selected portion. The following assumptions are taken while developing the model architecture. The caption either directly or indirectly explains the picture's theme. Captions provide more extensive data, not only about items and their features but also about occurrences, than typical picture annotation, which uses keywords to describe important objects. The picture's content is described in the document that comes with it. This is especially true in news reports because the photographs typically illustrate occurrences, objects, or persons referenced in the article. Images are not labeled fully. Some of the objects described in the image are not labeled therefore the paper assumes that information about all objects can be derived from the image-article pair.

Content Selection
The paper proposed a probability-based image annotated model. The model takes an assumption that photographs and their related text are predicted by a shared collection of latent parameters/topics. The paper describes the article and photograph by a simple multimodal dictionary comprising of words and ocular terms. Because of synonymy and polysemy, a large number of words in the dictionary points to the same concept. The model uses latent Dirichlet allocation (LDA) that is basically a text generating probability-based model. LDA expresses pictorial and textual essence combined as probability scattering over a collection of topics. The proposed model takes these topic scattering into consideration during discovering the most possible keywords for the photographs and its related article.

Extractive Caption Generation
The model uses an extractive caption predictor that is based on earlier work on mechanized summarization that mostly focuses on the extraction of sentences. The goal is to construct a synopsis by simply recognizing and appending the most essential sentences in a file. Abstracts for a wide range of documents, regardless of style, text format, or topic matter, may be created without a significant lot of linguistic study. For the paper, the model only extracts a single sentence. The main leading hypothesis the paper follows is that the sentence needs to be maximally equal to the expressive keywords predicted by the annotated model. Due to the likely nature of the method of the model, the paper is able to depict the content of the photographs in two

ways. First, as keywords list in a ranked manner. Second, distribution of topics.

The most basic way of calculating the similarity among keywords of photograph and article sentences is overlap word.

$$\text{Overlap}(W_I, S_d) = | W_I \cap S_d | / | W_I \cup S_d |$$

Here, W_I denotes the bundle of keywords predicted by the model and Sd denotes a sentence in the article. The chosen caption is the sentence with highest overlap with keywords of photographs.

The naive method of similarity computation is word overlap on the basis on lexical identity. This is overcome by denoting keywords and article's sentence in a vector space and calculating the similarity among the two vectors denoting the keywords of photographs and article sentences respectively. The method is to generate a word-sentence co-phenomenan matrix in which rows express the word and column express the sentence, and every entry the recurrence with which the word emerge within the sentence(taking the assumption that keyword are making a sentence).

The similarity of the vectors denoting the W_I keyword and S_d sentence of the article can be computed as cosine of angle between them.

$$\text{sim}(W_I , S_d) = W_I . S_d / | W_I | | S_d |$$

Abstractive Caption Generation
Although extractive approaches provide naturally grammatical templates with minimal language processing, there are some disadvantages to observe. As previously stated, there is frequently no single statement in the text that accurately defines the image's content. In most situations, the keywords may be discovered in the paper, although they are scattered throughout many phrases. Second, the chosen sentences result in large captions (often longer than the typical document phrase), which are not as brief and appealing as human-written captions. As a result of these considerations, the paper moves to caption creation and offer models based on single words/phrases.

6.3 Drawbacks

- Paper assumes the images are fully labelled. In this case the model is unable to find extract contextual information about the unlabelled object present in the image.
- Highest overlap is used for selecting the most appropriate sentence as the caption but it may happen that useful information is scattered along multiple sentence therefore, this model fails to generate useful caption.
- The use of LSTM in the model architecture increases the training time of the model. This results in additional time overhead.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- Named entities are the most important part of news image captioning and this paper doesn't deal with them, therefore, this paper is not good for news image captioning tasks.
- The model overperforms other baseline models but stills very far behind manual evaluations.
- Since the architecture is somewhat similar to encoder-decoder some important information is always lost midway in transitioning from high to low-level dimension vector.
- In LDA number of topics is fixed and cannot capture correlation therefore not a best strategy to use LDA in the model.
- A weight-based mechanism is needed to be present in assigning weight to each image and news representation as some articles may be more inclined towards text meaning rather than image representation.
- No image restoration techniques are incorporated in the model to take care of noisy and blurry images.

7 Paper 7 [7]

7.1 Importance

The value of news media cannot be overstated. These news outlets have a wealth of information hidden between the lines of their stories. Extracting information and structuring it in order to make conclusions is critical in analytics. The prime objective of the paper is to focus on building a machine to find details from news in the English language. Then this information is given to the user in an organized manner. The proposed mainly extracts the named entities like location, person name, organizations, the date that are present in the news article, summary, and important keywords related to the news.

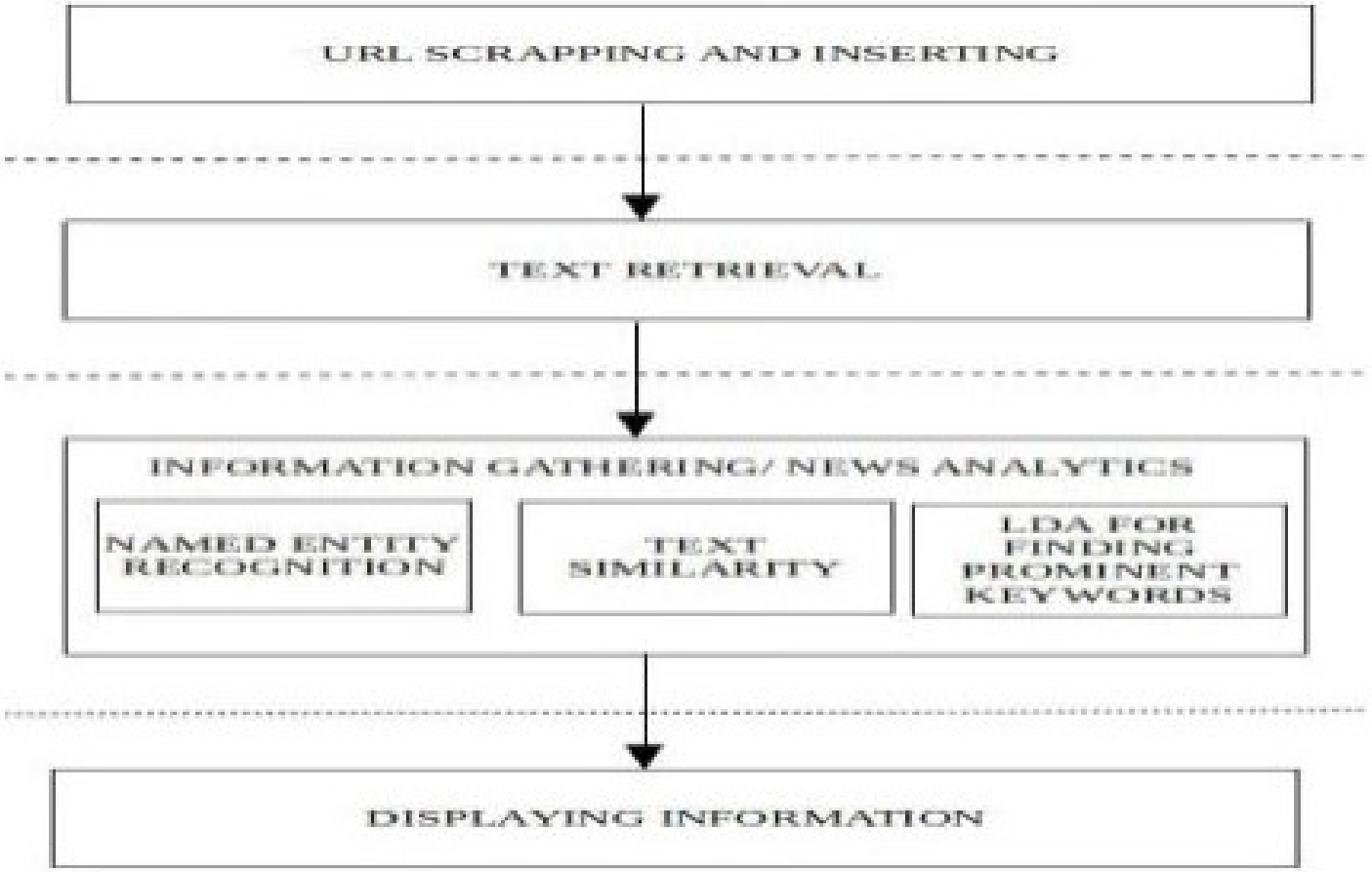

Figure 5: The sequence of components in the model.

The proposed model is deployed and tested for a few websites. There is a possibility for expansion of the developed tool. When expanding the machine, the arrangement and modularity would guarantee that there are fewer loopholes and that the modules function are not dependent on each other. Cleaning URLs, headlines, and dates, for example, occurs during the web scraping process. Then any date concerns won't have to be addressed at a later time. Additionally, the updated URLs make sure that the text retrieval step requires the addition of techniques to determine if the URL is legitimate or not legitimate.

More domains can be crawled with the program, and more data capture methods may be utilized to reduce the amount of material necessary to summarise the news. The tool's main goal is on Newspaper articles, but it may be extended to accommodate non-English-language content with appropriate translation systems. However, it is important to ensure that the text acquired is semantically accurate; otherwise, the accuracy of the facts gathered in succeeding phases will be compromised. It's also feasible if the mother language has been supported. The specifics can then be added to the structures for data shown in that language directly.

The only question is whether or if the application can be extended to store non-English elements. Another important aspect to notice is that if the web scraping component is enhanced, the tool may be used to collect data from other domains that are not mass media but still includes data (like blogs, Wikipedia). With keywords, a tiered search may be developed, and modules can be added (as long as retrieving and data collecting time is kept to a minimum and streamlined). The application can be extended to function as a micro search service.

The main distinction is that while data gathering components operate on the material, the specifics would be shown beside the URL. Memory, access time, and textual CPU utilization must all be handled in such instances. The concept of a search engine is simply a notion till then. The front-end may be changed in aside from working on the back-end. Tuning the design so that the user only sees content from a limited group of sites, which may be altered according to the customers' requirements, is one example. However, the article focuses mainly on the framework and presents a concept that may be used to accomplish a larger notion.

7.2 Methodology

The methods needs to be processed in the sequence as mention in the figure. The components are implemented in python language and database used is PostgreSQl. After every level, the output with get written in the database. The input for upcoming level will be extracted from the database. The execution of the output at every level before writing in the database, is compulsory to easy the upcoming process.

Before the actions can be done, the user must fill in the values in the media outlet and news domain. It is worth noting that the dashed lines in the illustration represent the independence of components from one another. This guarantees that the components may be extended without difficulty by the person operating the tool and that defects can be identified immediately. The independence also assures that if an error or anomaly occurs at any moment, the modifications performed at the modules may be undone without impacting the preceding or subsequent components in any manner. After resolving the issue and using the data gathered and saved from earlier components, the current component may be launched again.

URL Scrapping and Inserting

The URL of the article will be the first thing needed for accessing answers from traditional media. Web crawlers are programs that may be used for this purpose. The crawler will only operate for a restricted amount of domains due to its narrow capability. The URLs to articles or subsections/categories can be discovered on a website's home. The connections are identified by a specific set of tags within their HTML format. Scrappy is a web-crawling tool that may be used to extract content surrounded by a certain set of tags. Every domain uses a unique set of attributes to encrypt its link to a story.

This necessitates the use of the paper to create each component for every news portal. When the model runs the component, the crawler will begin with a collection of URLs from various websites (the URLs necessary to begin are also saved in the database). Crawlers will search for links to articles or other areas. It will retrieve the URL, date, and title for every article as well as seek for additional links on the site. If links are discovered, the crawler will go to the link. Crawlers will continuing to scan in this way until no more links are detected or already processed links are discovered.

Text Retrieval

Also, the text body of a news item is typically surrounded with website-specific tags (most notably the p> tag). To acquire material, web crawlers can be adapted and expanded. However, in such circumstances, the disturbance in the information would be significant. Another issue is that several news websites alter their encoding from time to time.

Crawlers must also be changed in order to obtain the required data. This is a demanding work, and the modules will be exposed to numerous changes on a regular basis. Newspaper3k is a Python package for parsing news items based on a URL. The library's methods may be used to retrieve the text from the URL. The function is straightforward to construct, there is less noise in the text, and it works regardless of how the article is represented. And that is why this component would demand less changes, particularly as the program is scaled up.

Information Gathering

The text and URLs collected in the preceding components are meaningless unless the data contained inside them is acquired. If the text is provided directly to the user, it is no better than the user seeing the articles straight from the website. This necessitates the need for knowledge extraction, which involves extracting relevant facts from the text, organising it, and presenting it to the user. The extracted data should display the characteristics from the material in a concise way while minimising detail loss.

Named entity identification

It focuses on identifying named entity keywords in the text article. This is done in a stage of three levels. The first level is to divide the article into tokens and relate every token to a speech. The second level is finding which token is related to the named entity. The third and final level is to find the entity type for the given entity on the basis of the speech tag part of the adjacent tokens. The model mainly deals with only three types of entities namely LOCATION, PERSON, and ORGANISATION.

LDA for finding keywords

The newspaper library includes features for locating keywords inside an article. However, the quantity of generated keywords may not always be adequate to provide context for the content. A customized version of LDA would be capable of extracting a notable collection of keywords. It is used to locate a topic for a writer. In general, the subject formed is a mixture of predominance for a collection of terms in the text. The topic would be assigned based on the article's most essential keywords. With the right rules, the algorithm may also be used to generate an impartial title for the material, assuming that the headlines are prejudiced but the content is not. Keywords are chosen from the most significant group of terms. The customized algorithm gives several topics, each a mixture of a given set of keywords.

Text Similarity

Article from different sites almost contains the same content. This will lead to redundancy in information. To solve this issue, the method needs to find text having similar meanings. The model uses the TF-IDFalgorithm that will first generate a matrix, with rows referring to articles and columns resembling to feature of the word. Every entry in the matrix will tell the weightage of a given word in a particular article. The weightage ranges from 0 to 1. Before the text is given input to the algorithm it has to be processed. Therefore the text is converted to lowercase and stopwords are eliminated along with the punctuations.

The cosine similarity approach, which analyses each article as a vector and determines the dot product for the set of vectors, is utilized. If the dot product of any two vectors exceeds a certain threshold, the articles matching to the = vectors will have almost identical content. The set of ids matching to the vectors is recorded and saved as parent id and child id, including the article's timestamp. The identical set of variables is saved in a database once more, with the only difference being that the parent and child ids are swapped. The parent, child value guarantees that just the contents of the parent article are shown, while the contents of the child articles are disregarded.

7.3 Drawbacks

- In LDA number of topics is fixed and cannot capture correlation therefore not a best strategy to use LDA in the model.
- Named entities are very important for the task of image news captioning the paper only deals with three type of named entities. Other named entities like DATE, QUANTITY, etc are not taken into consideration can result in a bad caption.
- No image restoration techniques are incorporated in the model to take care of noisy and blurry images.
- This model can't be used in certain fields as it is not able to give good results in the unorthodox news genre like satirical news, philosophical articles, etc.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- Not useful for conceptual abstract articles and news as the number of entities in such articles is very low.
- TF-IDF is based on the bag of the words model, therefore it cannot capture the static location in the text, semantics, co-occurrences in different articles, etc. Therefore, TF-IDF is only hepful as a lexical stage feature.
- TF-IDF algorithm Cannot capture semantics.
- In cosine similarity, the magnitude of vectors is not taken into account only direction is responsible for the final result.
- Not suitable for complex news where the content text is not directly related to the image.

8 Paper 8[8]

8.1 Importance

The work of creating news picture captions, on the other hand, is distinct from that of traditional image captioning. In contrast to standard image captioning tasks, where the input is merely an image, news image caption creation takes both a news storey and its associated picture as input. As a result, rather than outlining objects in a picture and explaining their characteristics or connections to one another, the output of news image caption creation is the enlightening text that not only describes the key linguistic features expressed in the given image, but also summarises the content of the corresponding news article.

A revolutionary deep NN-based framework or structure for automatic caption synthesis for photos of the news story is suggested in this study. The suggested methodology yields a superior BLEU score than variants of the models and behaves closely to the LDA strategy on METEOR scores, according to the experimental results of the model on the BBC News dataset. Nonetheless, human assessors preferred the captions created by the suggested technique above the captions provided by the LDA system the majority of the time, according to the report.

On the Web, there is a plethora of info. Several online media organizations, such as CNN, Yahoo, and the BBC, include photographs in their reports and even offer photo feeds connected to current affairs. These media sources are a great place to find multimedia files with data in the form of videos, photographs, and word-based writing. The availability of such a large volume of multimedia content has fueled the development of machine-learning-based systems that combine data from several perspectives. The goal of this paper is to provide predestined news images that capture all of the distinct items presented in the news story as well as their contextual connections.

8.2 Methodology

The paper follows a problem statement which is: having a news photograph I and its related document article D, generate a caption S that best express the photograph given D. Therefore, the training dataset comprises of a photo-article-caption tuple. In testing, the article and photo are given the paper should generate a descriptive caption.

The paper proposed a novel deep Neural architecture to predict the caption for the news photographs. The method first changes the sentences of the articles into a series of vectors with the help of a pre-trained embedding. The embedding used is order-embedding by Vendrov. The method then encrypts the related image into a representation of an image. Pre-trained Oxford VGGNet is used for image encoding as an off-the-shelf feature founder. After the encoding step, both representations are exposed to the same dimensional 1024 space. Then the series of vectors are directed to the LSTM cell network. The outcome of the network is again fed to another LSTM model cell that takes photograph representation as another input. The outcome of this LSTM cell is taken as a joint expression that expresses the contents of both photographs and their article. The objective for training the LSTM weights is the cross-entropy between the joint representation and order-embedding of the caption.

Image and Text Representation

For the purpose of text representation, the model uses a pre-trained order-embedding model by Vendrov. This embedding is based on the distributed representation. The main aim of the embedding is to leverage the partial order framework of the ocular-semantic hierarchy by understanding the relationship between the sentence and its semantic space. This results in a

1024-dimensional space.

For image representation, the model uses a pre-trained CNN. Complicated CNNs are consist of many cnn layers allowing them to learn complex and hidden features. Pre-trained Oxford VGGNet is used for image encoding as an off the shelf feature founder. The VGGNet network is made up of 22 layers. The model uses the fc7 fully connected layer features as the encoding of the photographs. The fc7 is the penultimate layer of the network. Then this encoding is projected to the dimensional space of the sentences. After this, both representation of the image and articles is in the same dimensional space that allows reasonable comparison among them.

Training the model RNNs are unquestionably effective in simulating patterns. However, the drawback of such networks is that they have been unable to convey data when the size of the chain exceeds a certain threshold. This is known as the vanishing gradient effect. To address this issue, an LSTM forgetting method has been developed. There are several LSTM variants. One cell is made up of three gates: input, output, and forget. Gates are often activated with sigmoid functions, whereas input and cell state are frequently converted with the tanh function.

An Long Short Term Memory has two input parameter at the timestamp t, one is xt the input vector at t, and ht-1 (hidden level vector at the t - 1 timestamp). W and b are trainable/configurable model parameters standing for weights and biases respectively.
In the forward pass, the updates to parameters are as follows:

$$
\begin{aligned}
\mathrm{I}_t &= \mathrm{g}(\ \mathrm{Weight}_{xi}\mathrm{x}_t + \mathrm{Weight}_{hi}\mathrm{h}_{t-1} + \mathrm{bias}_i\ ) \\
\mathrm{f}_t &= \mathrm{g}(\ \mathrm{Weight}_{xf}\mathrm{x}_t + \mathrm{Weight}hf\mathrm{h}_{t-1} + \mathrm{bias}_f\ ) \\
\mathrm{o}_t &= \mathrm{g}(\ \mathrm{Weight}_{xo}\mathrm{x}_t + \mathrm{Weight}_{ho}\mathrm{h}_{t-1} + \mathrm{bias}_o\ ) \\
\mathrm{cin}_t &= \mathrm{tanhactivation}(\ \mathrm{Weight}_{xc}\mathrm{x}_t + \mathrm{Weight}_{hc}\mathrm{h}_{t-1} + \mathrm{bias}_{cin}\ ) \\
\mathrm{c}_t &= \mathrm{f}_t\mathrm{c}_{t-1} + \mathrm{i}_t\mathrm{cin}_t \\
\mathrm{h}_t &= \mathrm{o}_t.\mathrm{tanhactivation}(\mathrm{c}_t)
\end{aligned}
$$

I_t depicts the input gate, f_t stands for forget gate, o_t resembles the output gate and cin_t is used for input transform.

Encoder-decoder models capture the information to a context vector, which is subsequently supplied to the decoder. During training, both phrase vectors and an image vector are given to an LSTM network in the forward pass to generate a context vector. The context vector is considered to summarise the information delivered in both written and visual representations. To backpropagate and adjust model parameters, we leverage the cross-entropy between the LSTM network output and the order-embedding vector of the picture caption as the loss function. We fix the learning rate calue to 0.6, the momentum value to 0.9, and train the model with sgd across 30 epochs. In testing, we recover the most appropriate sentence from a newspaper article depending on the cosine distance between the resultant vector from the LSTM and the order embedding vector of every sentence.

8.3 Drawbacks

- The complexities and subtleties of language, as well as the ethereal ephemeral nature of knowing something, make building a substantial knowledge base graph a challenging task (for example, there is a continuous evolution of knowledge and facts).
- There is a limit up to which we can take words/phrases with their key scores as input, increasing the number will exponentially increase the computing time of the overall model.
- The model is less scalable as the value of hyperparameters used in the model can impact the results very badly.
- There is also a need for some preprocessing methods for the articles. Punctuation, prepositions, adjectives, and other elements may be removed during preprocessing. The reason for this is that these words add nothing to the core text of the news story, and it will also make the process of feature representation via order embedding easier.
- No image restoration techniques are incorporated in the model to take care of noisy and blurry images.
- Not suitable for complex news where the content text is not directly related to the image.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- The unnecessary text associated with the article like repeating some sentences, associated notes, results in memory overhead.
- Semantic dimension reduction may result in loss of important description about the news photograph
- The use of LSTM in the model architecture increases the training

9 Paper 9 [9]

9.1 Importance

With the fast growth of Internet connectivity, readers' primary source of current events has shifted from conventional print media to the Web as a newscasts transporter. A typical Internet news article comprises of a narrative that summarises current affairs and a graphic representation of those happenings. Nevertheless, these network reports contain a high number of bogus stories with misaligned pictures and wording. It is simple to deceive consumers, disrupt the network ecosystem, and decrease the media's reputation if misleading news with misaligned graphics is not resolved quickly.

To address these issues, the study proposes a news detection scheme on the basis of the graph that can enhance the news abstract model and take advantage of it to compute the statistical correlation with the news text, allowing it to determine if news corresponds to the news picture misfit. The implementation of a news picture-text matching finding technique boosts social media's reputation and cleans up the network ecosystem. The current approaches have the accompanying two major drawbacks. First, by obtaining the fitting and comparing between picture description and textual ranking, a large number of linguistic mistakes will be created. Second, certain approaches use the Euclidean distance among image explanation characteristics and textual information as a criterion over whether or not they are fitting and thus overestimate syntactic closeness between image description and news narrative without taking into account their deep mutual information.

This work provides a way to enrich picture description semantics while reducing the grammatical structure and semantic intricacy of textual content in order to address the above two drawbacks. (1) We employ the concept-NET knowledge graph to augment news picture descriptions in order to reduce semantic errors between image descriptions and content. (2) To detect fake news with misfit text and text, we adopt a sentence similarity computation approach based on linguistic structure and interpretation to account the fitting problem between visual description and the text computation. The results of the experiments suggest that the system described in this research can successfully distinguish network news with misaligned images and text.

9.2 Methodology

The study proposes a technique for detecting news pictures and text matching. First, news text is subjected to text preprocessing step and semantic dimension minimizing , after which ConceptNet that is basically a type of an external knowledge graph, is used to enhance the picture overview of story, and eventually, a sentence closeness computation technique on grammatical structures is used to determine whether or not a news photograph is aligned.

Generating description

Given a story article with a photograph X as the input. O is a bundle of words derived from the news specimen X by "YOLO9000, O = YOLO(X)". The paper define o $\in$ O and I_o = ConceptNet(O). I_0 is denoting the representational term of "ConceptNet graph" of all the available components in news specimen X. i_o = ConceptNet(O) , i_o is a collection of linked words/phrases placed into ConceptNet by component O. O is a collection of words directly linked to news depicted by X. I is a set of words/phrases loosely related to news represented by X, as described in this study. The proposed method is used to compute the key of every phrase/word in collection I . The topmost three key words or phrases are used as input in this publication, with d representing vector representation of Encoding related specifically terms acquired by Word .n representing vector representation of loosely related terms/phrases receiving the top 3 essential scores of Word Representation.

a = ConvNN(X) , a represent the feature vector of news specimen obtained by the convolutional neural network. The model architecture takes a,n and d as the inputs to the language module. The language module is based on LSTM. The starting state of the language module is :

$$\text{x-1} = \text{a} \,||\, \text{d} \,||\, \text{n}$$

|| denotes the joint operation.

$$x_t = W_e S_t \ , \quad t \in (0, 1, 2, \ldots., \text{N-1})$$

The expression of S_t denotes the result at the time t - 1 that is being used as the input in next time t_1.

$$P_{t+1} = \text{LSTMModel}(x_t) \ , \quad t \in 0, 1, 2, \ldots., \text{N-1}$$

Resemblance Computation

When evaluating matching extent, the news text and pictures characterise the presence of sentence structure differences caused by miscalculation. This article uses the deriving semantic features in the assertion, the first to use gating convolution similar incomplete contextual features in system to retrieve sentences, by acquiring attention mechanism after big words in the sentence semantic associations, and then determine the resemblance between them.

1) Sentence input code

The paper take a assumption that the maximum size of the sentence is L. If the sentence's length is less than L than 0 is length is take to add to L . Otherwise the extra letter are discarded. S = [x_1, ..., x_L] , x_i is denoting the vector of word ith of the sentence S. The paper takes the mean of all vectors x_s in Sentence S as their sentence vectors and then combine the two parts

as the starting expression of the sentence, S' = [x_1, ..., x_L, x_s] as the start input code for the sentence. The x_s is computed as :

$$x_s = (1/L)\sum_{i=1}^{L} x_i$$

2) Sentence local semantic

In order to compute the local feature data of sentence, the study uses gated CNN to obtain the semantic data of sentences. Lets say, the sentence to be examined is S, the gated CNN used two liberated convolutional architecture with the same architecture, the one among the two is followed by the Sigmoid activation function and the remaining one is not followed by any activation. After this, the two are multiplied bit by bit to obtain the output result S' as shown below.

$$S' = conv_1(S) \text{ x } \sigma \text{ } (xonv_2(S))$$

Conv blocks with varying widths of conv kernel are used to convolute the sentence in order to better extract the location features of the sentence. Conv kernels Windows 1, 2, and 3 are used to convolute phrases, and every conv kernel has a number of 300. Then, to acquire the local semantic data description of sentences, the maximal pooling operation of length 3 is utilised, and the outcome of pooling three convolution units is combined.

This paper implements the definite integral procedure and the solace multiplication operation on the last semantic representation of two sentences, respectively, in order to compute the semantic relatedness between the news text and the depicted characterization, and then accomplishes the blending to obtain the similarity measure representation couple of the two sentences. To determine the semantic relatedness likelihood function of the two phrases, input the semantic relatedness of the sentences into the two complete conn layers, then use the Softmax algorithm.

$$Pairs = concatenation(|source + target|, source \text{ x } target)$$

Source denotes the text in the pairs of sentences and target denotes the news expressed characteristic sentence in the duel of sentences. + and x denotes the subtraction and multiplication of the related elements respectively. When there are 3 set of text utterances and photos to specify the resemblance is higher than 0.6, think this illustration matches news network news, and when there are 3 set of text utterance and pictures to define the resemblance is less than 0.6, think this illustration does not match the news.

9.3 Drawbacks

- The complexities and subtleties of language, as well as the ethereal ephemeral nature of knowing something, make building a substantial knowledge base graph a challenging task (for example, there is a continuous evolution of knowledge and facts).
- There is a limit up to which we can take words/phrases with their key scores as input, increasing the number will exponentially increase the computing time of the overall model.
- The model is less scalable as the value of hyperparameters used in the model can impact the results very badly.
- There is also a need for some preprocessing methods for the articles. Punctuation, prepositions, adjectives, and other elements may be removed during preprocessing. The reason for this is that these words add nothing to the core text of the news story, and it will also make the process of feature representation via order embedding easier.
- No image restoration techniques are incorporated in the model to take care of noisy and blurry images.
- Not suitable for complex news where the content text is not directly related to the image.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- The unnecessary text associated with the article like repeating some sentences, associated notes, results in memory overhead.
- The use of LSTM in the model architecture increases the training time of the model. This results in additional time overhead.
- Semantic dimension reduction may result in loss of important description about the news photograph.

10 Paper 10 [10]

10.1 Importance

The goal of news picture captioning is to dynamically produce subtitles or explanations for news photos. News picture captioning can generate draught captions for news photos, which is useful in applications since there are so many news pictures on the Web that manually creating captions is time-consuming and labor-intensive. Several digital news outlets, like CNN, BBC, and Yahoo!, include photographs in their reports and even offer photo streams connected to recent issues. These news websites are a valuable source of multimedia data including data in the form of video content, photos, and natural language texts, and hence offer annotated news image datasets for supervised captioning techniques.

News picture captioning differs from traditional image captioning in that news photographs are linked to news writing, hence captions for news photos should include data about the surrounding text. The caption of a news photograph frequently reflects the news story's unique occurrence. Generic image captioning, on the other hand, creates captions that only contain relevant data about pictures and cannot represent particular information in the news since it concentrates on the picture itself, which is the only information that can be employed to build a description. Because not all photos include linked words, such as news images, traditional captioning does not take into account related or neighboring text.

The fact that news photograph captions frequently include more depth data is one distinction. The news text contains all of the pertinent details. Another distinction is that news image captions frequently include data on particular events covered in the news. Because photos are generally embedded in the text of news, it's difficult to gain statements regarding event data alone from photographs. In the last few decades, deep learning-based picture captioning algorithms have taken important steps. These developed methods are mostly used to produce generic picture descriptions and rely on image data to do so. The supporting text of the images is ignored by most neural image captioning techniques.

By summarising the news text as per query image, this work provides a news image captioning approach that relies on the attentional encoder-decoder paradigm. To calculate the context vector, a multi-modal attentional technique is provided. The proposed method has been tested on DailyMail news picture captioning datasets, which are constructed by analyzing HTML formatted files and gathering photos, captions, and news texts. Experiments using the DailyMail test dataset reveal that the suggested strategy beats both conventional picture captioning and text analysis.

10.2 Methodology

The proposed model is multi-modal based on the attentional encoder-decoder architecture. It is used to build captions for images of news from the text article and news image. The reason for being multi-model is that the inputs of the architecture are news article and news image and the end product is a descriptive news image caption.

The model is primarily comprised of four parts: i) the encoder for text, ii) the encoder of the picture, iii) the decoder, and iv) the basic attention mechanism. The encoder for text is consisted of an RNN model to encode the article associated with the news. The encoder of the picture consists of a state-of-the-art CNN(VGGNet) model that gives a vectorial representation of the image. An RNN model works as a decoder to decode the input in form of words. The attention mechanism is used in this RNN model. The tuple $< I, T, Y >$ is used to represent image-text-caption information. T represents the text article, I represents the image, Y represents the news caption. GRU stands for the "gated recurrent unit" is taken as the RNN basic cell in the model architecture due to its efficiency and effectiveness as LSTM while low time consumption.

Encoders

The main objective of the text encoder is to encode articles and the image encoder has the objective of encoding the photos. The paper used a bi-directional RNN as a text encoder to encode associated article T of the news. T denotes a sequence of words. T is represented as (w_1, w_2, w_3, ..., w_T). W_i is denoting embedding of the i^{th} word calculated by the word2vec embeddings. |T| is the size of the T embedding. The vectorial expression of the T is calculated by adding the last hidden layers of the RNN.

The CNN model "Oxford VGGNet" is taken as the photo encoder to encode the photo into vectorial expressions. The VGGNet is earlier used in the classification of the images. It comprises many CNN layers. Every CNN layer is followed by a pooling layer(either max or min pooling). The last layer of the architecture is fcc layer. The last CNN layer divides the picture into 14 x 14 ocular parts. Evey ocular part is encoded into 512-dimensional vectorial expressions. The VGGNet is proved good for picture caption and for attention procedures to attend. Therefore, the model uses the last CNN layer as the vectorial representation of the photo. The photo I is denoted as (v_1, v_2, ..., v_{196}). Here, v_i is the CNN encoding of the i^{th} ocular component of the photo.

Decoder

The model uses RNN based decoder to decode the embedding representation of the article and news image to build a caption.

Equation (1): $\bar{h_0}$ = tanh(W^{dec0} x enc^{text}+ b)
Equation (2): h_i = GRU ^{dec1}($W^{softmax}$ $\bar{h_i}$ + b)
Equation (3): $\bar{h_i}$ = GRU^{dec2}(h_i,c_i)
Equation (4): $y_i \sim$ softmax($W^{softmax}$ $\bar{h_i}$ + b)

Equation (1) calcuate the earlier hidden state of decoder. Only text representation used in initial state calculation, the image one is used in attention procedure . Equation (1) and (4) represents the computation for decoder. The decoder used in the model architecture is based on duel-level hidden output. Equation (2) and (3) are used to output the next hidden level. hi is calculated from the former hidden level and the previously obtained word. The starting input is a token called <sos> which indicates the start of the sentence. Two procedure based on attention are proposed to calculate the ci by computing the dependence among the hidden decoding levels and text representation and image representations in the upcoming two subsections. Equation (4) takes the softmax to calculate the likelihood of every word from the hidden level. The decoding phase eliminates when it predict the word token <eos> which indicate the end of the sentence.

Mutli-Modal Attention

Conventional attention procedure computes the c_i by calculating the dependence among the hidden decoding levels and text representation and image representations for traditional image captioning during the calculation of the dependence among the hidden decoding level and text representation for text summarization. Images in news have both themselves and additional texts, therefore the contemporaneous attention procedure (AttSim) are proposed to compute the context that is based on both text and image representation in form of encoding.

$$\text{Equation (5): } h_j^{txt} = [\ h\rightarrow_j^{txt},\ h\leftarrow_j^{txt}\ ]$$
$$\text{Equation (6): } h_j^{img} = \tanh(\ M^{img}.\ v_j + b)$$
$$\text{Equation (7): } H^T = h_i^{txt} \mid 1<= i <= |T|$$
$$\text{Equation (8): } H^I = h_j^{img} \mid 1<= j <= 196$$

Equation (5) to (8) depict the text and picture representation. Equation (5) calculates the hidden level during the encoding of word j^{th} by adding the forward hidden level $h\rightarrow_j^{txt}$ and backward hidden level $h\leftarrow_j^{txt}$. Equation (6) calculates the hidden level of the pictorial section j^{th} by modifying the dimension of v_j with M^{img}. This is then succeeded by tanh activation. The size of M^{img} is found to be $|h_j^{txt}| \times 512$. $|h_j^{txt}|$ denotes the dimension of h_j^{txt} . Equation (7) represent H^T as the bundle of the hidden text levels. Equation (8) represent H^I as the bundle of the hidden image levels. The multi-modal attentional procedure calculates the scores of attention by attending the text and image representation synchronously. The multi-modal synchronous attention procedure broaden conventional attentional procedure from applying attention to text representation only or image only to apply attention to both of them in form of two parallel paths.

10.3 Drawbacks

- VGGNet contains a large number of trainable parameters therefore the training is very slow.
- Taking the overall image representation instead of the concerned object in the image(central object and avoiding background information) is not a way of encoding the images for news image captioning tasks.
- Named entities are the most important part of news image captioning and this paper doesn't deal with them, therefore, this paper is not good for news image captioning tasks.
- Model's performance decrease over a large number of entities in the sentence or for long sentences as the attention mechanism is not effective for longer text.
- The use of GRU incorporates slow convergence and low learning efficiency in the model architecture.
- No inter-text semantic preserving encoding is used therefore the model can't tell similarity between synonyms before the training starts.
- The unnecessary text associated with the article like repeating some sentences, associated notes, results in memory overhead.
- Not suitable for complex news where the content text is not directly related to the image.
- Since the architecture is somewhat similar to encoder-decoder some important information is always lost midway in transitioning from high to low-level dimension vector.
- This model can't be used in certain fields as it is not able to give good results in the unorthodox news genre like satirical news, philosophical articles, etc.

11 Comparison

11.1 METEOR

'Metric for Evaluation of Translation with Explicit Ordering' is a full form of METEOR. It is a measure for evaluating the performance of machine translation tasks. The computation of this metric is done by taking the harmonic average of unigram precision and also recall which takes priority on precision. Together with the basic precise word similarity, it also offers other characteristics not present in other measures, such as stemming and synonymy comparison. The measure was created to address some of the shortcomings of the more common BLEU statistic while simultaneously producing a strong correlation with human judgment at the phrase or segment level. The BLEU measure, on the other side, seeks connection at the corpus scale.

11.2 BLEU

BLEU stands for bilingual evaluation understudy. It is a method for assessing the appropriateness of model-translated text from one natural speech to another speech. The connection between a model's result and that of a person is believed to be quality: "the nearer a machine translation is to a skilled human interpretation, the superior it is" - this is the core notion underlying BLEU. BLEU became one of the first metrics to promise a strong correlation with human quality judgments, and it is still one of the most common computerized and low-cost measures.

Individual translated sections, often sentences, are scored by matching them to a set of 'high-quality reference translations'. These are then aggregated over the whole dataset to check the overall quality of the translation. Pronunciation and grammatical accuracy are not considered in evaluating the bleu scores. Bleu score ranges from 0 to 1. Values closer to 0 means it is less related and values closer to 1 means it's more

11.3 CIDEr

Consensus-based Image Description Evaluation is a full form of CIDEr. It is used to evaluate photo descriptions using human consensus. The main goal is to compute for image I_i how well a candidate sentence c_i is similar to the consensus of the set of photo descriptions. The words present in the reference and candidate sentences are mapped to their root forms in the preprocessing step. For example, fish, fishing, etc all get reduced to fish. The TD-IDF algorithm is used for the method for calculating the CIDEr score for the sets of n-grams for every sentence. The overall CIDEr score is the average of all n-gram individual CIDEr scores. The primary goal of the CIDEr score is to evaluate translations of model architecture based on human resemblance without having to make arbitrary decisions on balancing content, grammar, saliency, and so on concerning each other.

11.4 Embedding

Images consist of several pixels and are represented in the form of a matrix of RGB values. For an image of the size of 1024*1024, there are a total of 1024*1024*3 numerical values and for a dataset of 10000 images, the total numerical value depicting an image rises to a level of 10000*1024*1024*3. Such a large amount of data is not practical for training a model therefore the image must be reduced to a simpler dimension that eases the burden of the model architecture and ensures fast obtain results. Similarly, in the case of the text, the machine is not able to take values other than numerical therefore the text must be changed to a numerical value sequence. This sequence should preserve the semantic and lingual features of the language.

Therefore to solve the aforementioned challenges embeddings are used. In the context of NNs, embeddings are low-dimensional, trained continuous vector representations of discrete variables such as text, image, etc. For encoding the images CNN is used that consists of several convolutional and pooling layers that reduce the overall dimensional while keeping the information intact and the last fully connected layer maps these reduced values to generate the final encoded representations. In the case of text, mapping-based embedding is used where each text in the dictionary is mapped to discrete numerical values. These values when stacked form a vector embedding that resembles the sentence whose words constitute individual numerical value. Some complex text embedding instead of taking individual mapping encodes the entire sentence into a vector using highly sophisticated algorithms.

For image representation usually, the pre-trained VGG19 embedding is used. The last softmax layer is not taken into account. For text, Word2vec and glove embeddings are the most common methods. However, when both image and text are extracted into the same semantic dimension thus there is a need for general embedding that can be used for both modalities. For example, order-embedding. The embedding acts as input features for the model architecture and is used for predicting the final caption on the basis of generated numerical vectors of the model. In cosine similarity, the resultant vector is compared with the embedding of the sentences of the article to obtain the most suitable sentence as a caption.

11.5 CNN

A Convolutional neural network is made up of one input layer, one output layer and a series of hidden layers which are basically a group of convolutional layers. It is a type of artificial neural network. It is widely used in tasks that deal with two dimensional data like images. However it can also work with data that is of one or three dimensions. ReLU is used as activation function and pooling, fully connected and normalization layers are also used in the network.

Convolution is the operation that is performed by convolutional layers. Convolution is basically an operation that is linear in nature. It multiplies the input with an array of weights. In case of input being of two dimensions as in images, the array of weight is also two dimensional and is known as kernel or a filter. The size of the filter is less than the input and dot product is done between filter and the input data.Dot product is basically an element wise multiplication. The logic behind using the filter that is smaller than the input is that the filter can be repeatedly applied on the input array a number of times at different locations of the input. In practice the filter is applied on the input array over different segments that have the same size as that of the filter from left to right and also top to bottom. The multiplication outputs a single value and doing it several times over different regions of image gives an array that has two dimensions. This output array of two dimensions is called a feature map.The idea of applying the filter on the image in this systematic way is quite powerful and it allows a filter designed to identify a particular feature to find that feature present at any location in the image. In the process of training the filters are optimized by the model via automated learning which can then produce meaningful feature maps that contain specific features

to fulfill the aim of the neural network which can classification of images(for instance).

CNNs are a type of multilayer perceptron which means networks that are fully connected. Every neuron present in a layer is connected with all of the neurons present in the next layer. In the case of CNNs the connectivity at this level may lead to overfitting. In order to solve the issue of overfitting weight decay is used during training, also dropout layers are added. Convolutional layers take inspiration from biological processes as the connectivity among the neurons represent the structure of the visual cortex of animals. Single neurons are responsible for stimuli in a limited area of the visual field which are known as receptive fields. The whole visual field is then covered by overlapping receptive fields of neurons. There are several applications of CNN in fields like image recognition, image captioning, video recognition, recommender system, image classification, analysis of medical images etc.

11.6 Named Entity

News articles heavily consist of named entities. Therefore the news image captioning task must incorporate the entities inside the generated descriptive caption. A named entity is a real-world element for example a person's name, location, date, organization, etc. Traditional image captioning doesn't take named entities into consideration therefore can't be used in captioning news articles. One of the main challenges in news image captioning is to train the model parameters as the named entities are very low in number and often are specific to the news articles. Therefore in training, the model should be able to handle any new entity that the model did not encounter before. This results in implementing new entity identification techniques in the model. One of the best techniques is to use the industrial-grade named entity identification tool SpaCy. SpaCy is pre-trained to predict the occurrence of the named entities in the article. SpaCy categorizes the named entities into different categories such as PERSON, PLACE, ORGANISATION, DATE, QUANTITY, etc.

After the named entities are identified the next big challenge is to predict a caption that is grammatically correct and lingual features preserved in it. There the three main ways of doing this. The first is to use cosine similarity and then pick the most suitable sentence having named entities as the caption. This method ensures the grammar is taken care of but fails to predict accurate captions as the useful information is often scattered along with multiple sentences. The second method is to use a placeholder and then generate the template and fill it with named entities and the last method is to predict the caption word by word. The last method is very impractical as the vocabulary must contain all the named entities beforehand and the model should have encountered multiple instances of such entities in training. The other problem that makes the task even more challenging is to deal with the captions that are metaphorical, ironical, or sarcastic in nature. For example, the bald eagle is a term used for the USA, and five-star flag is used for China, etc.

11.7 LSTM

LSTM's full form is Long short-term memory and it is a type of Recurrent neural networks. Whenever a human reads an article, he makes up the context and meaning not on the basis of reading single words only; rather he does so on the basis of the words he read earlier in the sentence. Feedforward neural networks have no way of classifying the event going on in a clip. It has no provision of using the previous information to make sense of the current information being portrayed. RNNS have feedback loops as opposed to feedforward networks which aid in these types of tasks. RNNs face the problem of dealing with dependencies that are long-term. In other words, in the process of predicting a word on the basis of previous information, if the gap is huge between the information and the place where we are predicting the word, RNNs can't learn to map this information in order to predict the current word.

Fortunately LSTM was designed to solve this issue of dependencies that are long-term and hence it has the ability to learn long term dependencies.They can process data points like images and also process data in sequential form like speech. Recurrent neural networks are made up of a series of repeated neural network components. This component that is repeating is simple in nature, for instance it can be one layer of tanh. Like RNNs, LSTM also has this chain structure, the difference being that the repeating component is not made of a single layer, rather it has 4 layers.These 4 layers interact with each other in a unique manner.

A typical unit of LSTM is made up of 4 components: 1 input gate; 1 output gate; 1 forget gate and 1 cell. The cell state acts as a conveyor belt to manage the information flow without much changes and is the key to LSTM.It remembers the value over different time periods. The process of adding or removing information from the cell state is carried out carefully by the gates.The gates act as a means to allow information to go through selectively. Gates consist of a sigmoid layer along with a pointwise multiplication operation. Sigmoid layer gives output that is greater than or equal to 0 and less than or equal to 1. Depending on the output of the sigmoid layer, gates know the amount of information that needs to be passed through. An output of 0 signals that no information should be passed through while an output of 1 signals to pass through all information. LSTMs have several applications like predicting on the basis of time series data, learning rhythms, detection of homology of protein, learning grammar, speech recognition, translation of sign language, forecasting traffic in short term etc.

11.8 Cosine Similarity

The closeness of two vectors in an inner product space is computed by cosine similarity. It is measured by taking the cosine of the angle among two vectors and checking if two vectors are heading in the same general direction. It is frequently used in text studies to analyze article similarities. In news article captioning, the cosine similarity is often used to obtain the most appropriate sentence in the article. This sentence in most cases is used as a caption. This method is used when the model instead of generating caption word by word follows an approach of closeness to named entity or keywords. Therefore this

method is grammatically correct but has a disadvantage in some cases. Another model where the cosine similarity is used is the probabilistic model based on the informative keywords.

The cosine similarity arranges the article's sentences in the order from most appropriate caption to least. This arrangement eliminates the need of associating weights to the individual sentences and the top K sentences can be used in text representation for feeding in the main model architecture. Such an approach eliminates the useless sentence in the article and thus reduces the dataset size. The reduced size with only suitable sentences allows the model to predict the best results in the least time. Tweaking the value of K can further elevate the quality of the captions.

In most cases, the value is K is fixed but in advanced cases, the sentences up to some fixed value of cosine similarity are taken. Cosine similarity is also used in comparing the textual and pictorial representation if the dimensional space of both representations is the same. Cosine similarity can also be used in considering only the most suitable objects of the image in image encoding rather than taking all the objects into account for the representation purpose. In filling the placeholder with entities, cosine similarity is used for choosing the most suitable named entity for the placeholder of the drafted caption.

The formula of cosine similarity:

$$\text{Cosinesimilarity(a,b)} = \cos(\theta) = \text{a . b / (| a | | b |)}$$

11.9 Transformer

Transformer network is a model belonging to deep learning that uses self attention mechanisms in order to account for the importance of different segments of input data. Primary uses of this network lie in the domain of NLP and computer vision tasks. Just like RNNs it works with sequential data for translations and summarization tasks as well.Difference being that it does not necessarily process data in sequential order.It figures out the context that provides meaning to words. Transformer has an architecture of encoder decoder. Encoder is responsible for processing the input data from one encoding layer to another. Similarly, the decoder uses the output of the encoder and processes it using decoding layers. The encoder has two parts: feed forward network and a self attention mechanism.

The input is processed by the first encoder by taking input as the embedding of the sequence of input along with positional information which is important to use the order present in the sequence. The output of this encoding layer is then taken through a self attention mechanism which judges the significance of input encodings to output next encoding.The feed forward network then does processing of these encodings one by one which are then passed to another encoder and to the decoder as well.

The decoder has three parts: self-attention mechanism , attention for encodings, feed forward network. It works quite like the encoder with the difference that an extra attention is applied over the encodings generated by the encoder to get relevant information. Positional information and output sequence embedding is taken as input by the first decoder. After the final decoder there is a linear transformation layer and a softmax layer which is responsible for predicting the probabilities of different possible words.

In every layer of the transformer network there is presence of multi-head attention. A single attention head is responsible for attending to tokens relevant to that particular one. With multi-head attention it can be done for different meanings related to relevance. For instance, the attention head can attend verbs to objects and thus encode the relations of relevance which makes sense. Computation is done for every head parallely which makes the processing faster and the output of attention layers are joined and then feed forwarded to the feed-forward network.

11.10 Object Detection

Object detection refers to the task of identifying different types of objects that are present in a given image. The task of identifying the class of a single object present in the image is known as image classification. Input for this task is an image with utmost one object and output is the class of the object(image in this case) .Pointing out the location of different objects present in the image by outlining them with bounding boxes is called object localization.Input for this task is an image containing objects that can be greater than or equal to 1.The output of this task is a tuple of bounding boxes where each bounding box refers to the location of an object present in the image.

When we merge these two components to get different objects with bounding boxes and the class that they belong to is called object detection. Input is an image containing objects greater than or equal to 1. The output is a tuple of bounding boxes for each object along with a label that refers to the class of the object. The class can be chair, person, cycle etc. In practice object recognition means object detection. For the task of object detection, R-CNN has been proposed. The latest technique belonging to this family is Faster-RCNN.
Faster-RCNN has two components:

- Region Proposal Network: It puts the feature map provided by a deep CNN into a neural network to get bounding boxes and also class prediction for different objects.It behaves like an attention mechanism for the next component.
- Fast RCNN : It extracts features from the regions that were proposed by RPN and finally gives the bounding boxes and classes of objects as the output.

Also YOLO (you only look once) has been proposed as well as another model for the task of object detection. The methodology is quite simple for this model; it first divides the image into multiple cells in the form of grids. Each cell has the responsibility of outputting bounding boxes with coordinates and also the confidence. Then the cells with the same classes (which have higher confidence scores) are combined to get final bounding boxes and class labels for objects present in the image.

YOLO9000 was put forward which has the capability of detecting 9000 different types of objects. It uses bounding boxes with already defined shapes and sizes of bounding boxes which are adjusted in the training step. Object detection has a lot of applications in fields like facial recognition, pedestrian detection, counting vehicles, activity recognition, tracking objects like football during a football match etc.

11.11 Comparison Table

The comparsion table is as follows:

Features/papers	1	2	3	4	5	6	7	8	9	10
LSTM	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗
CNN	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
Object Detection	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
Transformer	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
Embedding	Byte-Pair	LGBPHS	-	CBOW	Word2Vec	-	-	Order	Word2Vec	Word2Vec
BLEU	18.10	-	-	9.53	19.6	-	-	34.27	78.3	7.24
METEOR	8.27	-	-	4.78	5.2	-	-	6.77	27.1	16.68
CIDEr	43.01	-	-	14.05	-	-	-	-	1.169	-
Named Entity	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
Cosine Similarity	✗	✗	✓	✓	✗	✓	✓	✓	✗	✗

12 Summary

News image captioning is a subset of conventional image captioning. In the new multimedia age, there is an ongoing competition in delivering news first to the consumer. Therefore News image captioning becomes a most important task. An ideal research paper should include some important features namely named entity, object recognition-based image encoding, and word-to-word generation. Papers surveyed here doesn't fully incorporate all the features. for example, paper 1 includes a method concerning named entities that are very specifics to this task but fails to capture only useful objects in the image, and similarly Paper 2 deals with the object recognition but fails to capture named entities. A total of 10 feature-based comparisons is done in the survey ranging from overall architecture to evaluation metrics(results).

The survey covers all the important aspects of the news image captioning task. The most important feature is named entity recognition a total of 4 papers deals with this feature. This shows that named entity is a very new and emerging feature in the computer vision field. Word-to-word caption generation is another important thing an ideal paper should consider. Most of the paper in the survey dealing with named entities fails to generate word-to-word captions. Instead of this, they capture the most similar caption out of the sentences of the articles on the basis of cosine similarity. Paper 1 is the best research study published among all the other papers mention in this survey. Paper 1 also takes account in learning hidden patterns of the long sentences. The attention mechanism incorporated in this paper along with the sequence-based generation of captions is giving very good results in comparison to other papers.

News image captioning can further be categorized into different fields like sports news, satirical news, checking fake news, etc. The papers surveyed deals with a different aspect of the task therefore the values of evaluation metrics differ drastically. Apart from this, caption generation for satirical news is a very challenging task and none of the papers mentioned can effectively deal with satirical news. The reason is the ambiguous relationship among named entities and the different interpretation of their meaning in different cases.

This survey presents a review on various types of techniques of image captioning and also provides a clear view of applications used in news image captioning. The survey also discusses the pros and cons of each paper briefly and provides a list of features that an ideal paper should consider. At last, all the paper mentioned are compared on some important features like embedding, object recognition, evaluation metrics etc.

References

[1] Z. Yang and N. Okazaki, "Image caption generation for news articles," pp. 1941–1951, 01 2020.

[2] X. Su and H. Zhou, "Automatic focus personage identification in multi-lingual news image," pp. 64–69, 10 2017.

[3] Z. Yumeng, Y. Jing, G. Shuo, and L. Limin, "News image-text matching with news knowledge graph," *IEEE Access*, vol. PP, pp. 1–1, 07 2021.

[4] Y. Jing, X. ZhiWei, and G. GuangLai, "Context-driven image caption with global semantic relations of the named entities," *IEEE Access*, vol. 8, pp. 1–1, 01 2020.

[5] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "Breakingnews: Article annotation by image and text processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 03 2016.

[6] Y. Feng and M. Lapata, "Automatic caption generation for news images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, 05 2012.

[7] G. Karumudi, R. Sathyajit, and S. Harikumar, "Information retrieval and processing system for news articles in english," pp. 79–85, 11 2019.

[8] V. Batra, Y. He, and G. Vogiatzis, "Neural caption generation for news images," 2018.

[9] S. Gao, H. Xu, Z. Chen, and Y. Wang, "Detection mechanism of news-text matching based on knowledge graph," pp. 264–267, 01 2021.

[10] J. Chen and H. Zhuge, "News image captioning based on text summarization using image as query," pp. 123–126, 09 2019.