TRAITEMENT AUTOMATIQUE DES LANGUES AVANCE

Master Informatique

2ème Année - 1er Semestre

Intervenants CM:

Gaël DIAS (gael.dias@unicaen.fr), Marc SPANIOL, Fabrice MAUREL,

Navneet AGARWAL, Kirill MILINTSEVICH

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Plan de l'UE

1. **[CM 1]** Représentation sémantique de texte [GD]
2. **[CM 2]** Cohérence textuelle [MS]
3. **[CM 3]** Modélisation thématique [NA]
4. **[CM 4]** Résumé de textes et traduction automatique [MS]
5. **[CM 5]** Génération langagière I [KM]
6. **[CM 6]** Génération langagière II [KM]
7. **[CM 7]** Sentiment Analysis, Image Captioning [NA]
8. **[CM 8]** TAL et web [MS]
9. **[CM 9]** TAL et handicap visuel [FM]
10. **[CM 10]** TAL et psychiatrie [GD]

11. **[TP 1-5]** Génération neuronal de comptes-rendus médicaux [NA - KM]

Sentiment Analysis

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

Sentiment Analysis

Sentiment analysis (often refered to as opinion mining) is the process of gathering and analyzing people's opinions, thoughts and impressions regarding various topics, products, subjects and services.

- Rapid growth of internet based applications such as social media and blogs.
- People's opinions can be beneficial for
  - Corporations: product reviews on amazon, sentiment towards a certain feature or service.
  - Governments: public opinion on policies, expected voting tendencies before election.
  - Entertainment industry: movie reviews, game reviews.
  - Travel: reviews of restaurants, hotels, tourist attractions, etc.

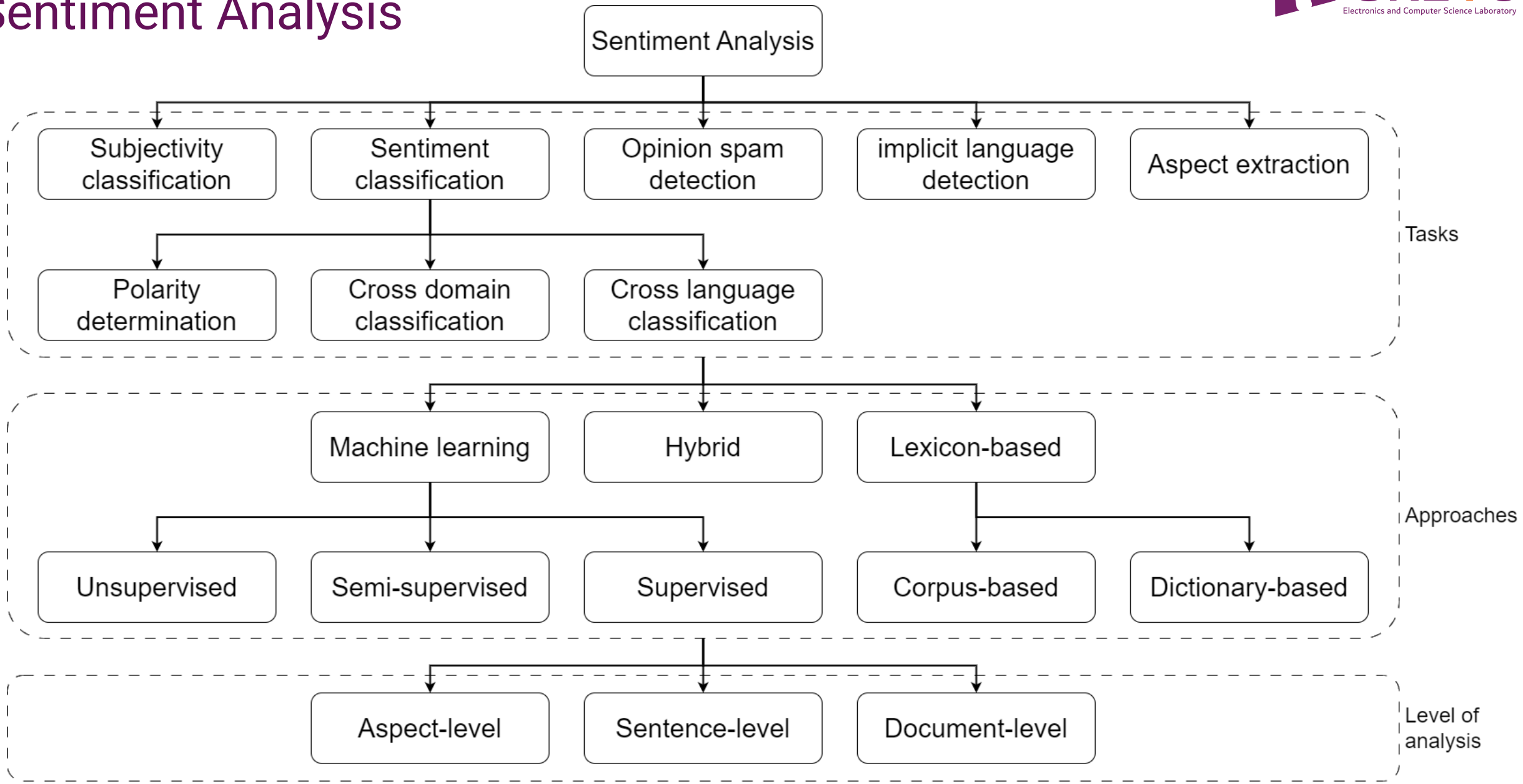
Sentiment Analysis

Sentiment analysis is a broad concept that consists of many different tasks, approaches and types of analysis.

Cambri at al. argue that a holistic approach is required , and only classification or categorization is not sufficient. They present it as a 3 layered problem that includes 15 NLP problems:

- Syntactic layer: Microtext normalization, sentence boundary disambiguation, POS tagging, text chunking and lemmatization.
- Semantics layer: Word sense disambiguation, concept extraction, named entity recognition, anaphora resoluion and subjectivity detection.
- Pragmatics layer: Personality recognition, sarcasm detection, metaphor understanding, aspect extraction and polarity detection.

Sentiment Analysis



Tasks

❑ Sentiment classification

- Most widely known and researched task in sentiment analysis.
- It can be divided into three major sub-tasks: polarity classification, cross-domain classification and cross-language classification.
- Polarity is usually classified as positive or negative with some researchers including a third category neutral.
- Cross-domain classification models transfer knowledge learned from data-rich source domain to a target domain where data and\or labels are limited.
  - Extract domain invarient features whose distribution in the source domain is close to that in target domain
- Cross-language classification fulfills the same function but for languages.
  - An example can be to train the model in source language with abundant data and testing it on target language by translating the input to source language.

Tasks

- Subjectivity classification
  - The goal of subjectivity classifiation is to restrict unwanted objective data objects for further processing.
  - It detects subjective clues, words that carry emotion or subjective notion like ‘expensive’, ‘easy’, ‘better’, etc.
  - These clues are used to classify objects as subjective or objective.

Ce livre coûte 10 euros → Cannot be used for sentiment analysis

Ce livre est cher → Can be used for sentiment analysis

Tasks

❑ Opinion spam detection

- Opinion spams refer to fake or false reviews intelligently written to either promote or discredit a product.
- Three main features are considered within this context:
  - Review content: the actual text of the review
  - Meta-data: information like IP address, geo-location, user-id, etc.
  - Real-life knowledge: this meathod utilizes learned experiences to classify spam. For example, if a aproduct has good reputation and suddenly inferior ratings are given over a period, reviews of that period might be suspected.

Tasks

❑ Implicit language detection

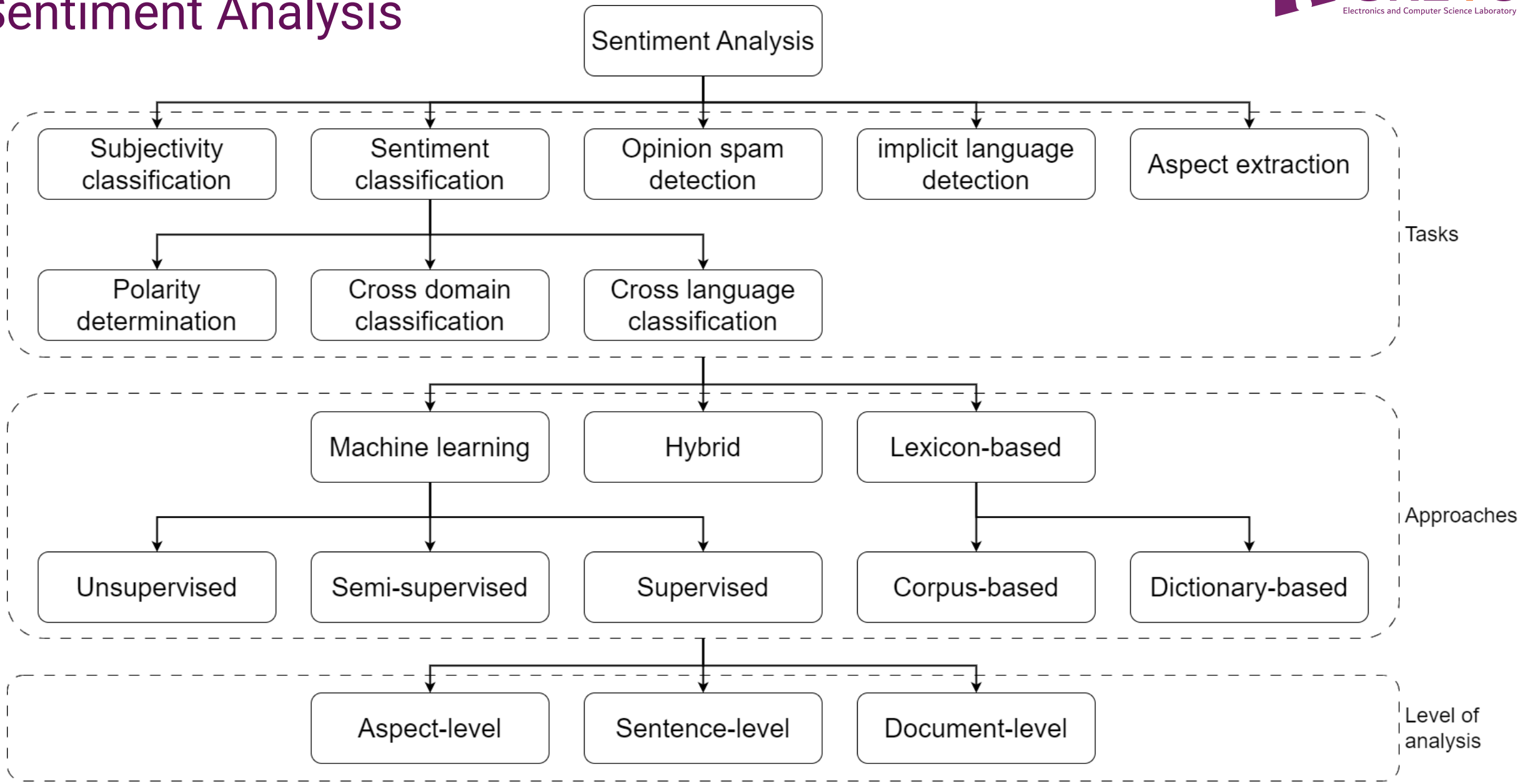
- Implicit language includes humor, sarcasm ans irony.
- There is vagueness and ambiguity in this form of speech, which is sometimes hard to detect.
- An implicit meaning can sometimes completly flip the polarity of a sentence.
- Example, « I love pain », pain is a factual word with negative polarity. The contradiction between pain and love can indicate sarcasm.
- Traditional methods for detection include exploring emoticons, expressions of laughter and heavy punctuation mark usage.

Tasks

❑ Aspect extraction

- Aspect extraction refers to retriving the target entity and the aspects of the target entity in the document. The target entity can be a person, product, event, etc.
- Aspect extraction is particularly important in sentiment analysis of social media and blogs that often do not have predefined topics.
- Most traditional method for this is frequency based where most frequent nouns and compound nouns are considered as candidates for aspects.
  - Not all nouns are aspects
- Syntax-based methods find aspects by means of syntactic relations they are in. For example, identifying aspects that are preceded by a modifying adjective that is a sentiment word.
  - Many relations need to be found for complete coverage

Sentiment Analysis



Lexicon based approaches

- Traditional approach for sentiment analysis that scans through the text for words that express positive or negative feelings to humans.
- It shows to be extremly dependent on domain of interest due to differences in language usage between domains.
- There are two main approaches to creating sentiment lexicons: dictionary based and corpus based.
- Dictionary based appraches start with an initial list of terms and iteratively expand the lexicon by adding synonyms and antonyms of the the terms to the list.
- They work best for general purpose use.
- Corpus based approaches starts with general purpose list of words and finds others terms from domain specific copus based on co-occurring word patterns.

Machine learning based approaches

- Machine learning based approaches can be divided into three categories: unsupervised, semi-supervised and supervised.
- Unsupervised approaches group unlabelled data into groups based on similarity to each other.
- Semi-supervised methods use both labelled and unlabelled data in training process.
- In cross-domain or cross-language classification, domain invarient features can be learned with unlabelled data and then labelled data can be used for fine-tuning the model.
- Supervised learning usually provides the best performance but also requires significantly more human effort in order to generate the labels.
- Machine learning approaches are especially popular for aspect extraction task with topic modelling being the most commonly used approach.

Image Captioning

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

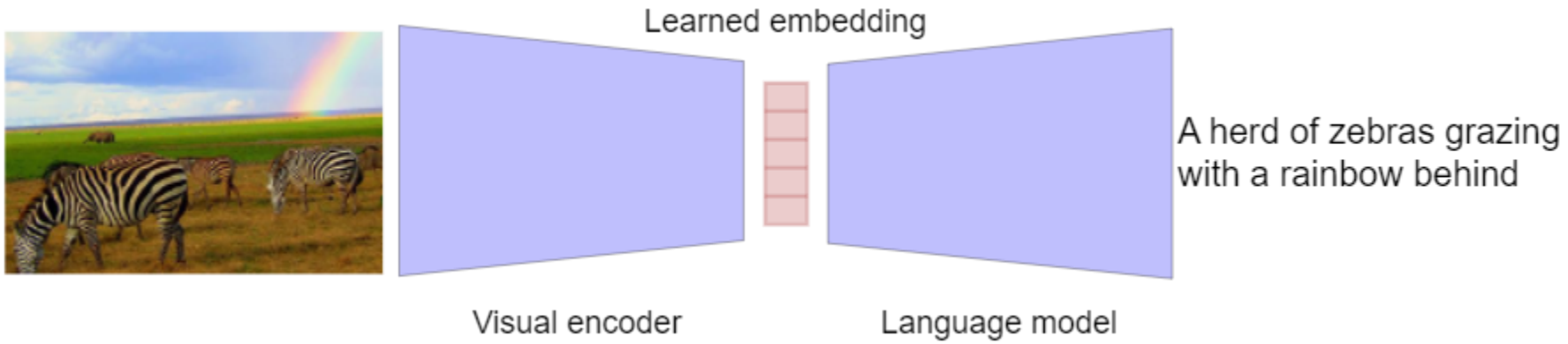
cnrs

Image captioning

Image captioning is the task of describing the visual content of an image in natural language.

- Visual component: model for understanding the visual data.
- Language component: model for generating meaningful and syntactically correct text based on learned image representation.

In its standard configuration the task is a image-to-sequence problem whose inputs are pixels and output is text.



Visual Encoding

Providing an effective encoding of the visual content is the first task within image captioning pipeline.

- Non-attentive methods based on global CNN features
- Additive attention methods
  - Grid based
  - Region based
- Graph-based methods
- Self-attentive methods employing transformer based paradigms.
  - Region based
  - Patch based
  - Image-text early fusion

Global CNN features

With advent of CNNs, all models consuming visual inputs have been improved in terms of performance

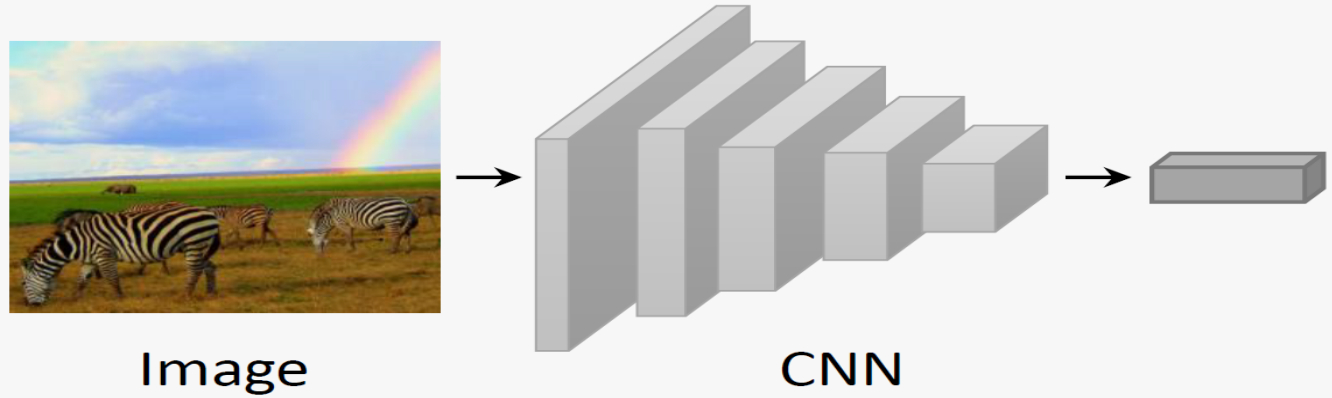
In the most simple recipe, the activation of one of the last layers of a CNN is employed to extract high-level representations, which are then used within language models for generating final text.

Advantages

- Simplicity
- Compactness of representations

Disadvantages

- Excessive compression of information
- Lack of granularity
- Unable to produce specific and fine-grained descriptions

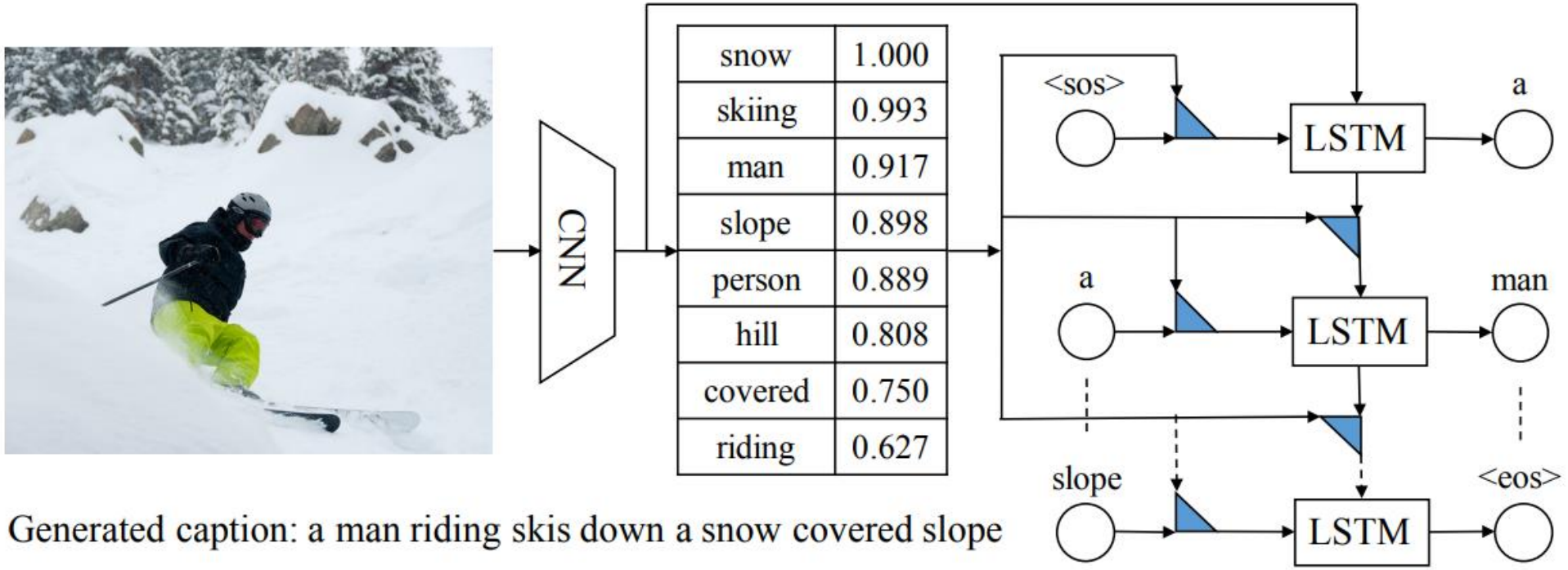


GoogleNet
AlexNet
VGG

Global CNN features

Gan et al., CVPR 2017

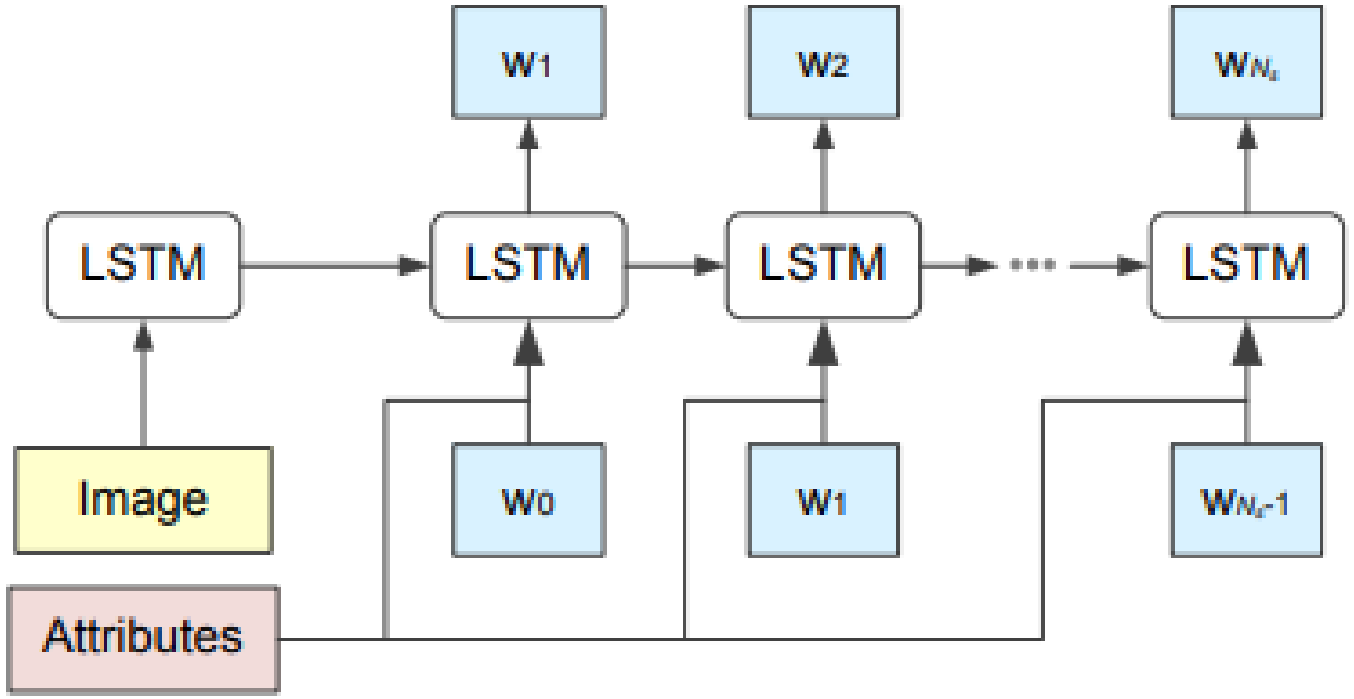
- ResNet-152 pre-trained on imagenet dataset



Global CNN features

Yao et al., Boosting image captioning with attributes, ICCV 2017

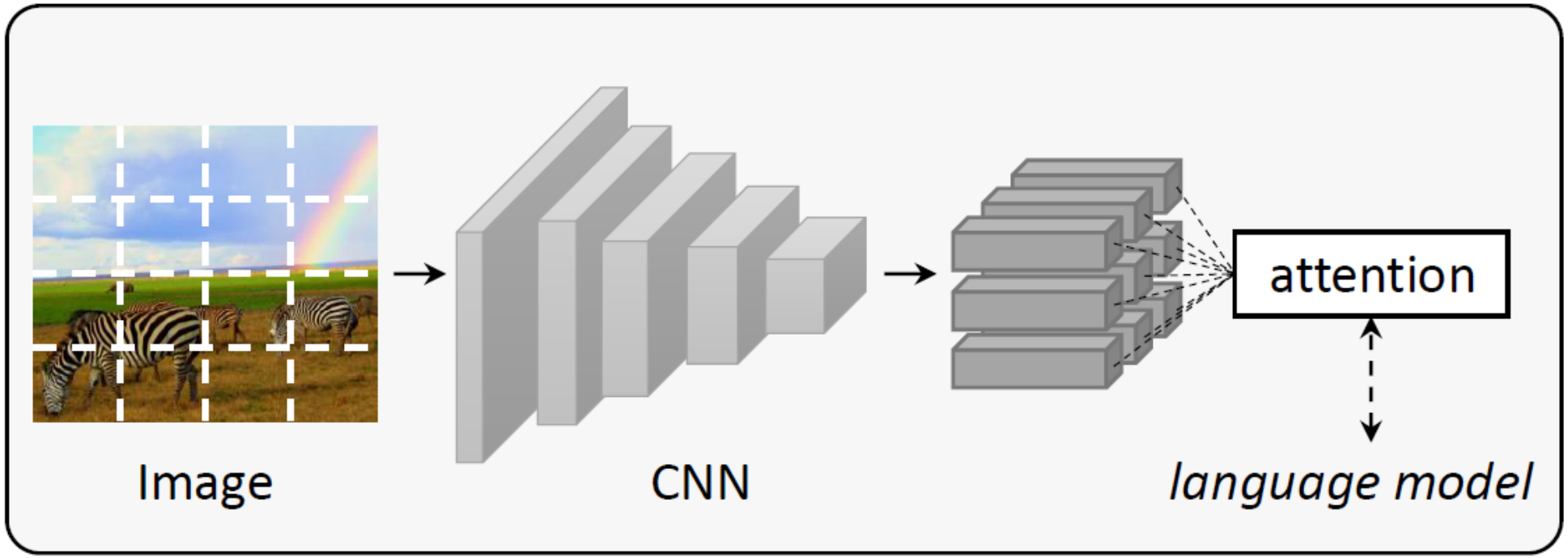
- 1,000 most common words on COCO as the high-level attributes and train the attribute detectors with MIL model (Fang et al., 2015)
- GoogleNet for image encoding



Attention based models

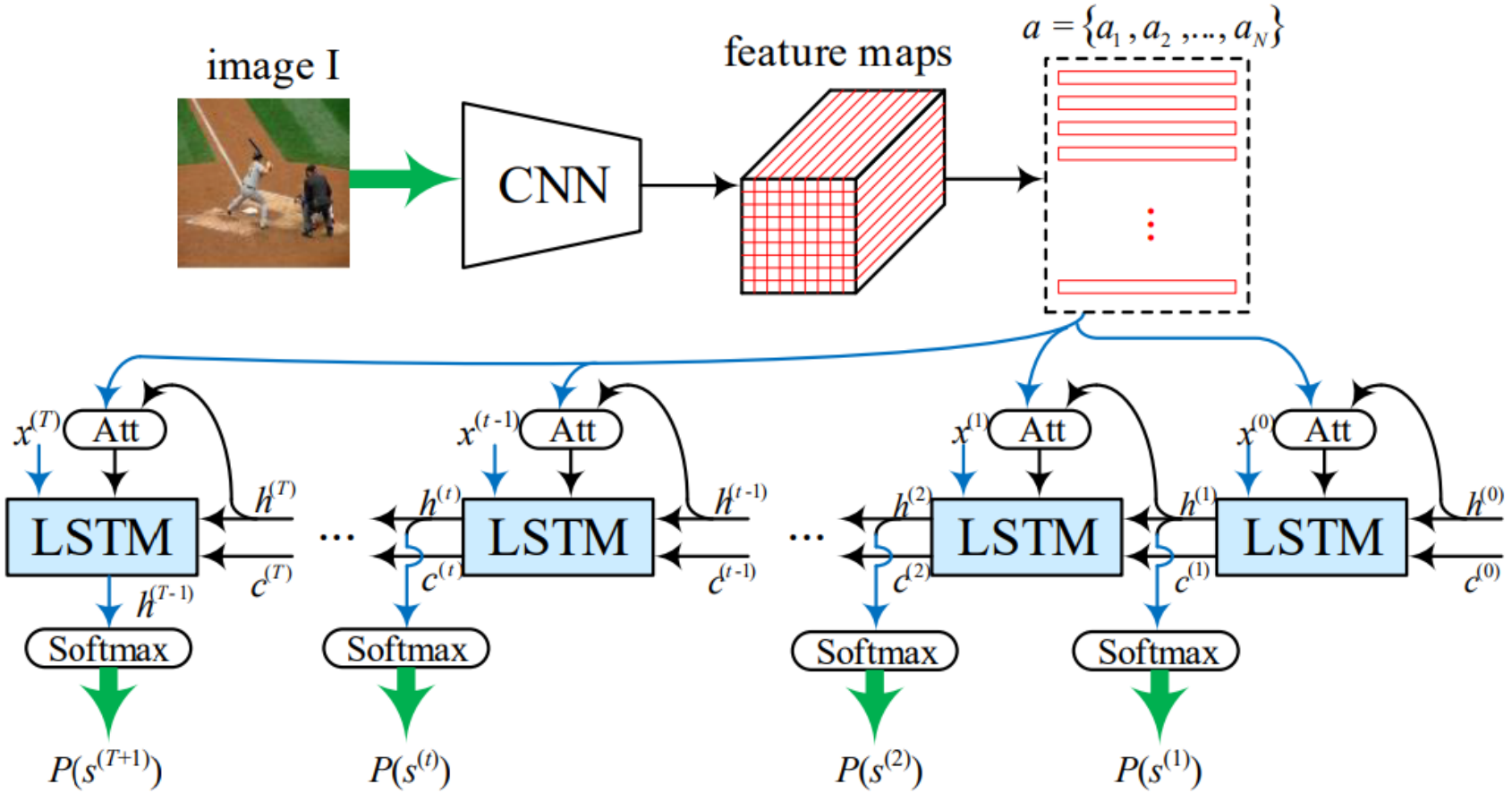
Grid based attention

- Improves upon the drawbacks of global representations and increases the granularity level of visual encoding
- Motivated from use of attention in machine translation.



Attention based models

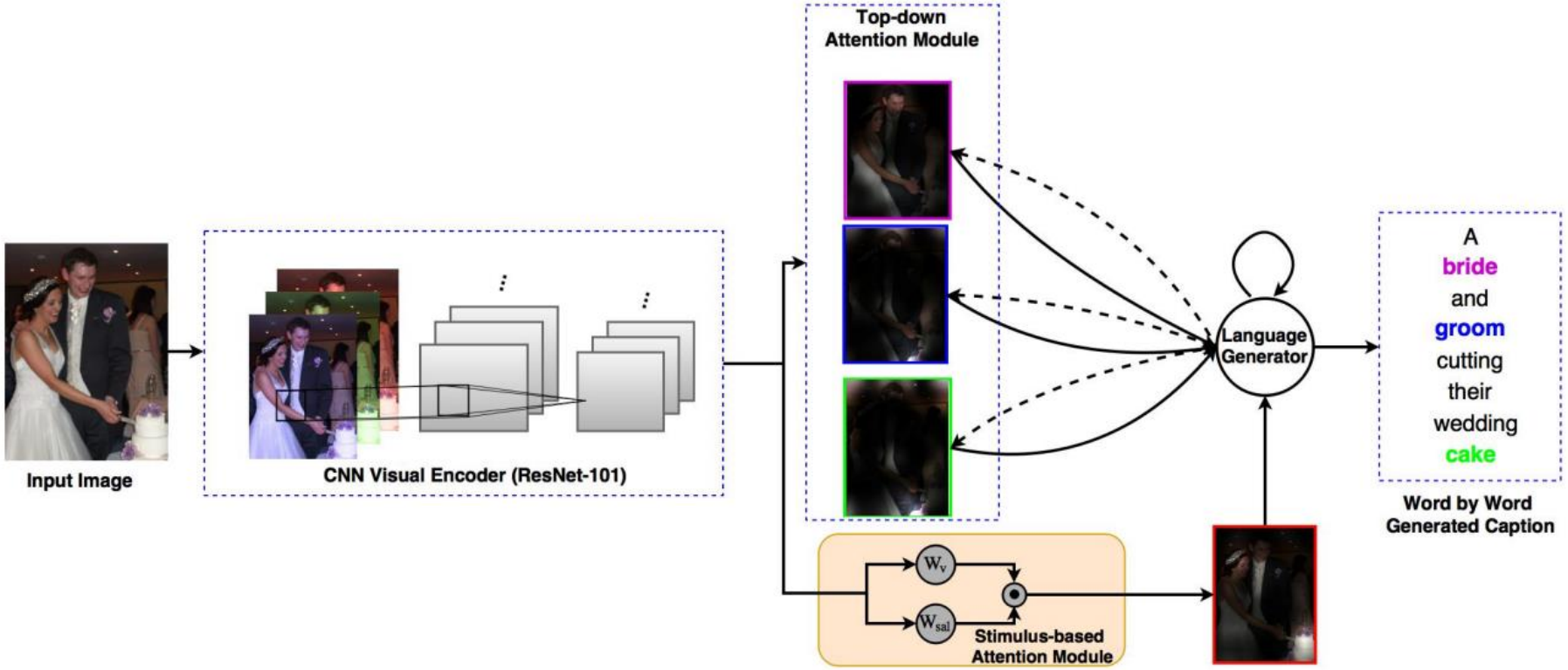
Ge et al., Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style, ICCV 2019



Attention based models

Chen et al., Boosted attention: leveraging human attention for image captioning, ECCV, 2018

- Improve the attention mechanism by incorporating human attention input into the model.



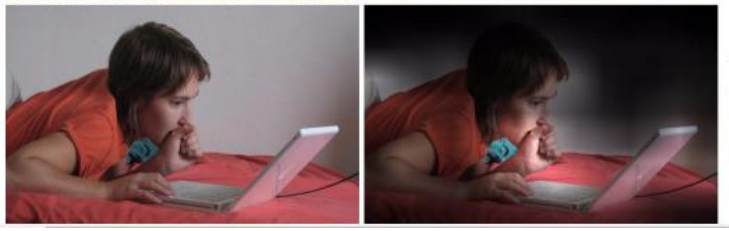
Attention based models

a parking meter on a street with palm trees.
a street sign on the side of the road.
1. A close up of a crosswalk sign in the middle of the road.
2. A tatter street sign sits in the crosswalk.
3. The yield to pedestrians sign is all scratched up.

a green and yellow bus parked next to a street sign.
a sign is on the side of a road
1. A green sign says Thruway one fourth mile.
2. A road sign stands next to the road.
3. a street sign below a bunch of power lines

a plate of food that is sitting on a table.
a bird sitting on top of a plate of food.
1. A plate topped with bread, greens and pasta and a bird.
2. there are two birds standing on the plate of food.
3. A bird attempting to bite a piece of sandwich bread.

a man is standing next to a motorcycle.
a man riding a motorcycle with a mountain in the background.
1. A man in a red shirt and a red hat is on a motorcycle on a hill side.
2. A man riding on the back of a motorcycle.
3. Man riding a motor bike on a dirt road on the countryside.

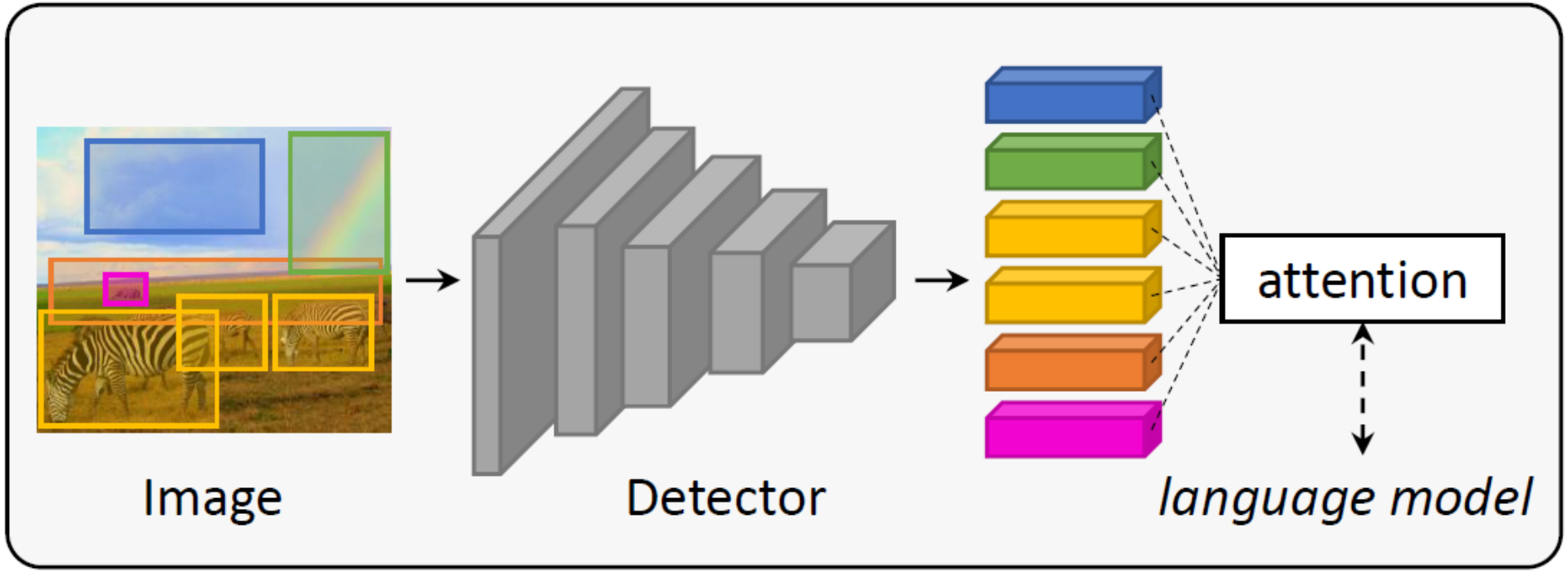


a woman sitting on a bed with a red shirt.
a woman sitting on a bed with a laptop.
1. there is a woman laying in a bed using a lap top.
2. A girl on a bed studying something on her laptop.
3. a woman using a white laptop on the bed.

Attention based models

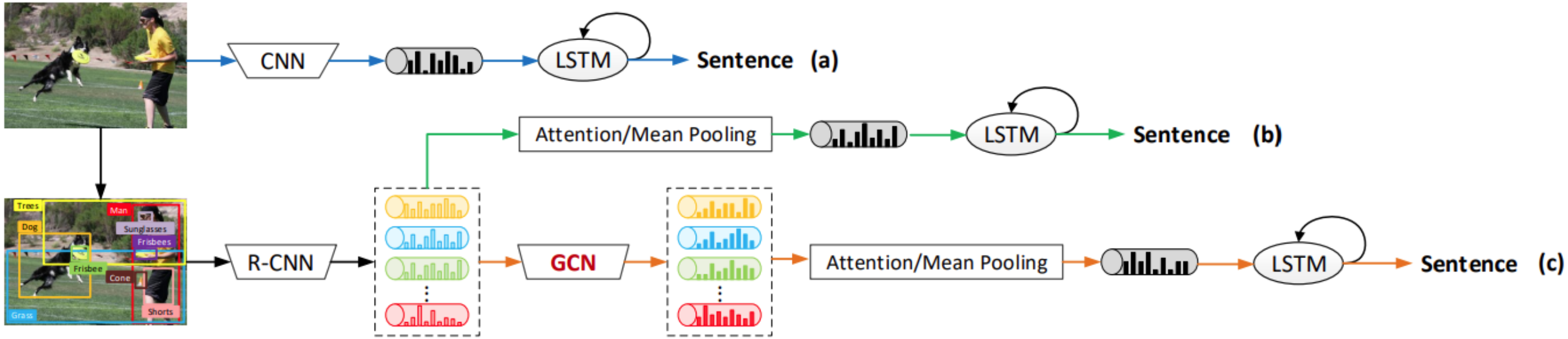
Region based attention

- A natural division of image is not in terms of grid but rather into regions.
- This is evident from saliency maps obtained for human vision.
- Region based attention models divide the image into logical regions rather than grids and calculate attention over these regions.



Graph based models

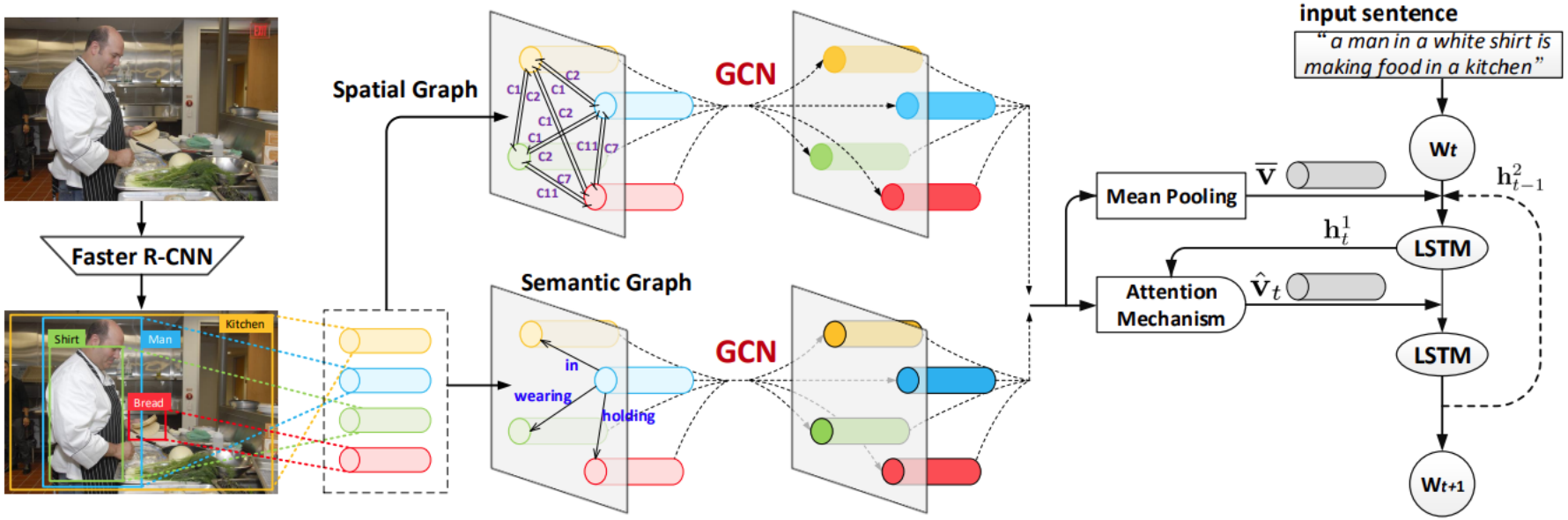
- Region based attention treats all regions equally without taking into account the interactions between them.
- Some studies consider using graphs based models for improved encoding of image regions by incorporating relations between different regions.



Graph based models

Spatial and semantic graphs

- A semantic relation classifier is trained on a large corpus and directly used for semantic graph generation.
- Spatial graphs are built and assigned depending on their Intersection over Union (IoU), relative distance and angle.



Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 684-699).

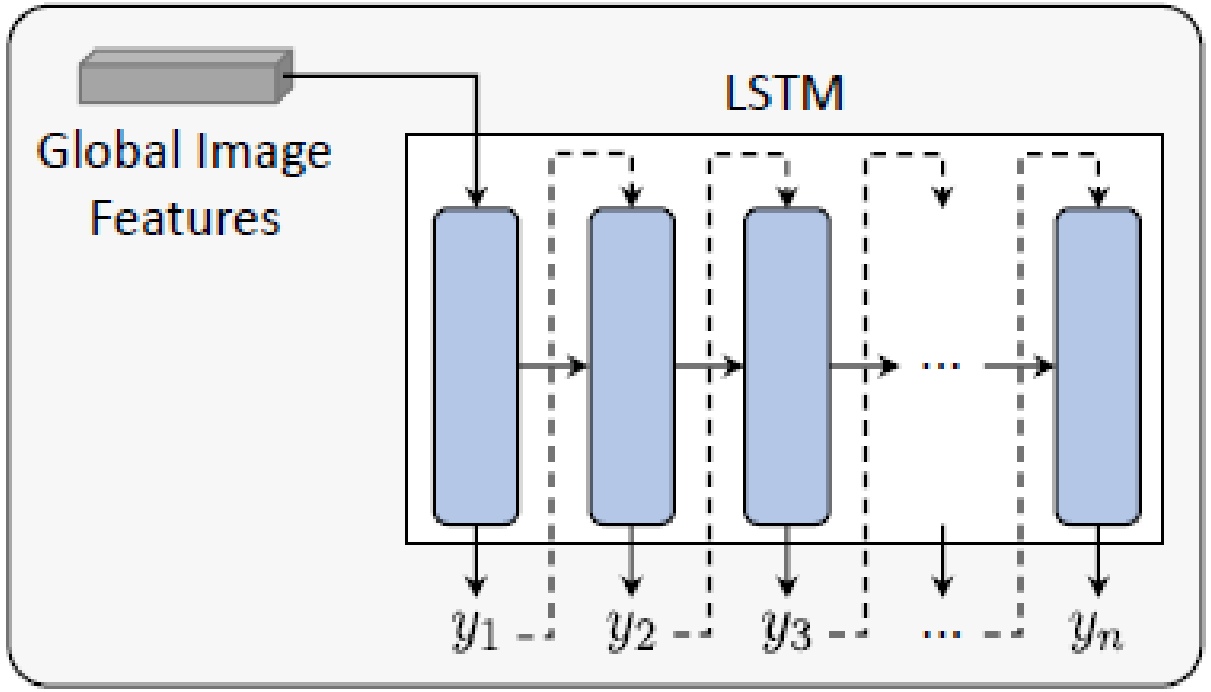
Language models

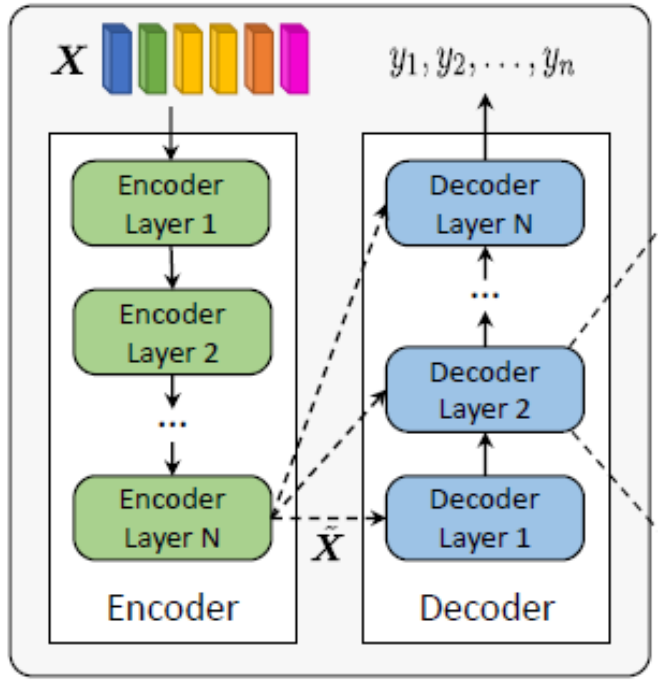
- Given the learned representation of the image, language models generate textual that is meaningful and syntactically correct.
- The goal of a language model is to predict the probability of a given sequence of words to occur in a sentence.
- The model makes incremental predictions where probability of i^{th} word in the sequence is conditioned on the preceding sequence.

$$P\left(y_1, y_2, \ldots y_n \mid \boldsymbol{X}\right) = \prod_{i=1}^{n} P\left(y_i \mid y_1, y_2, \ldots, y_{i-1}, \boldsymbol{X}\right)$$

Language models

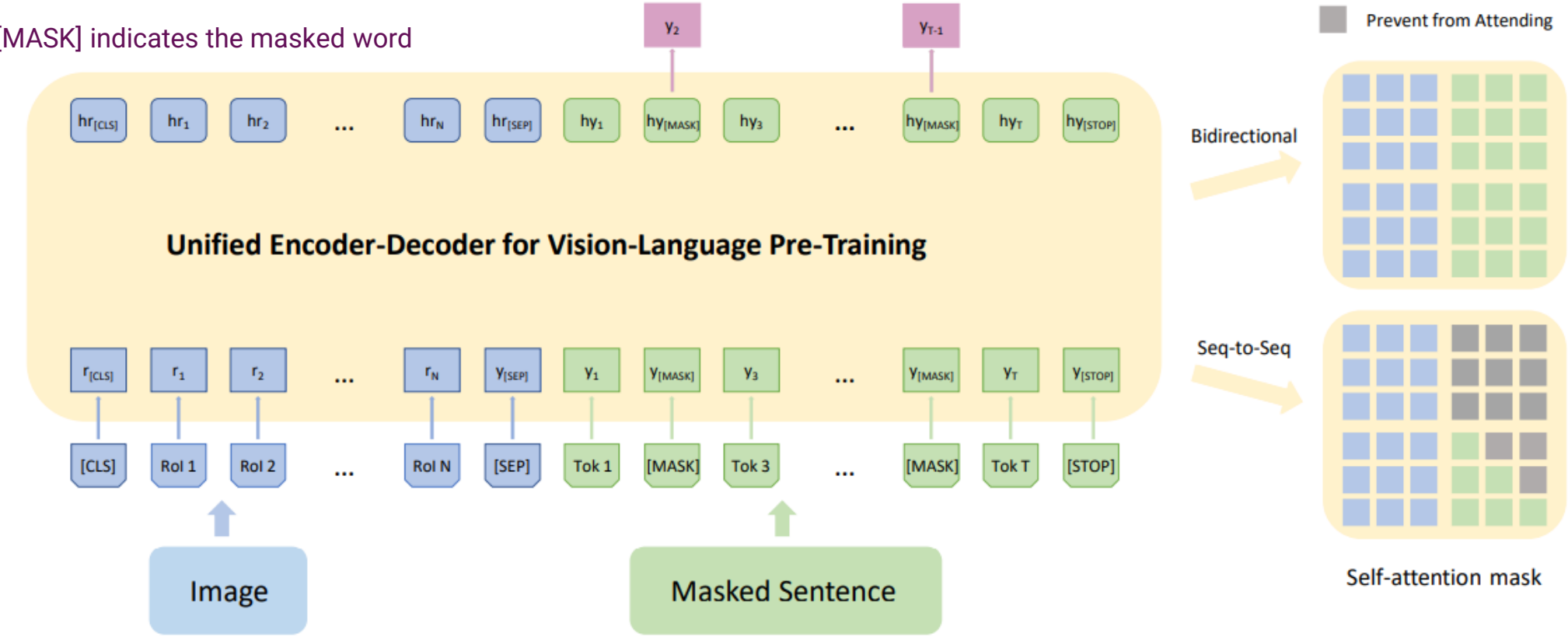
- The main language modeling strategies applied to image captioning are:
  1. LSTM based
  2. Transformer based fully attentive approaches
  3. Image-text early fusion (BERT-like)





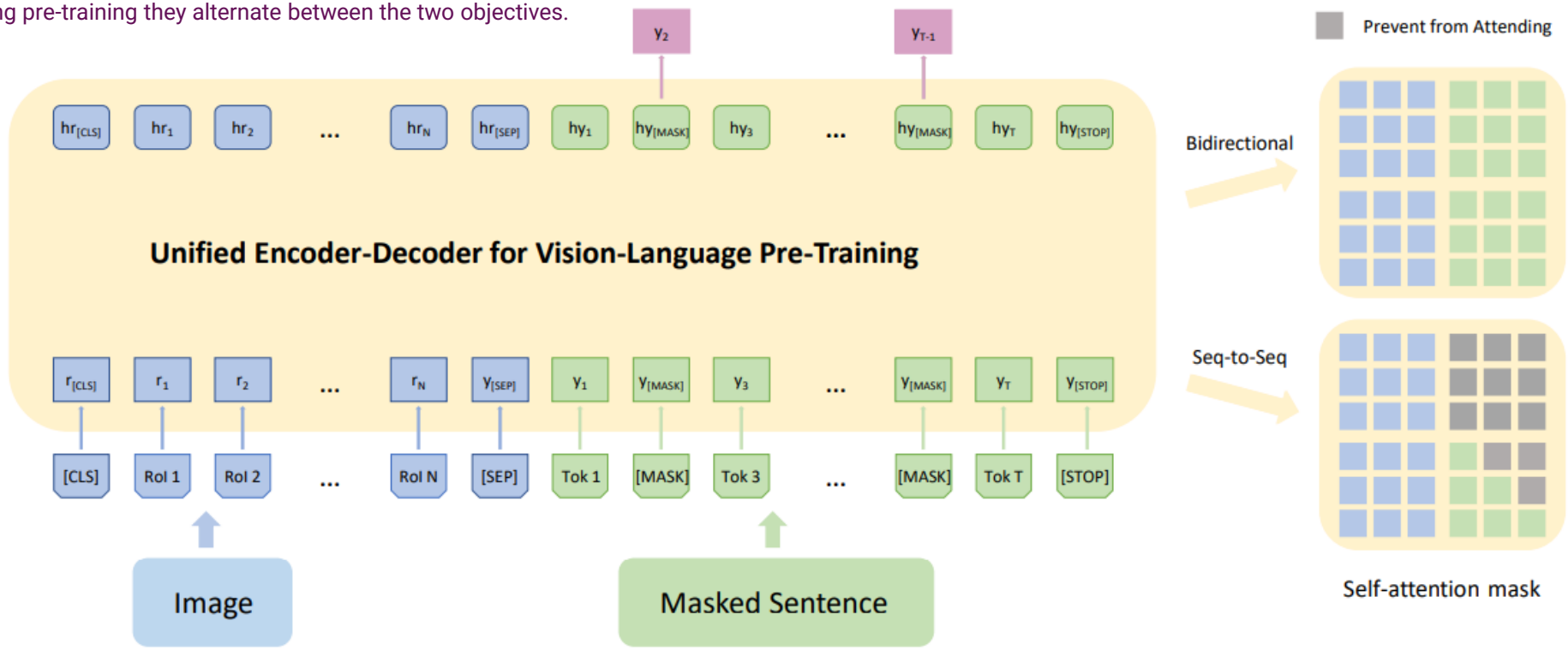
Vision-Language Transformer Model

- Unifies the transformer encoder and decoder into a single model.
- Model input consists of class-aware region embeddings, word embeddings and three special tokens [CLS], [SEP] and [STOP].
- [CLS] indicates the start of the visual input.
- [SEP] indicates the seperation betzeen visual and text input.
- [STOP] indicates end of sentence.
- [MASK] indicates the masked word



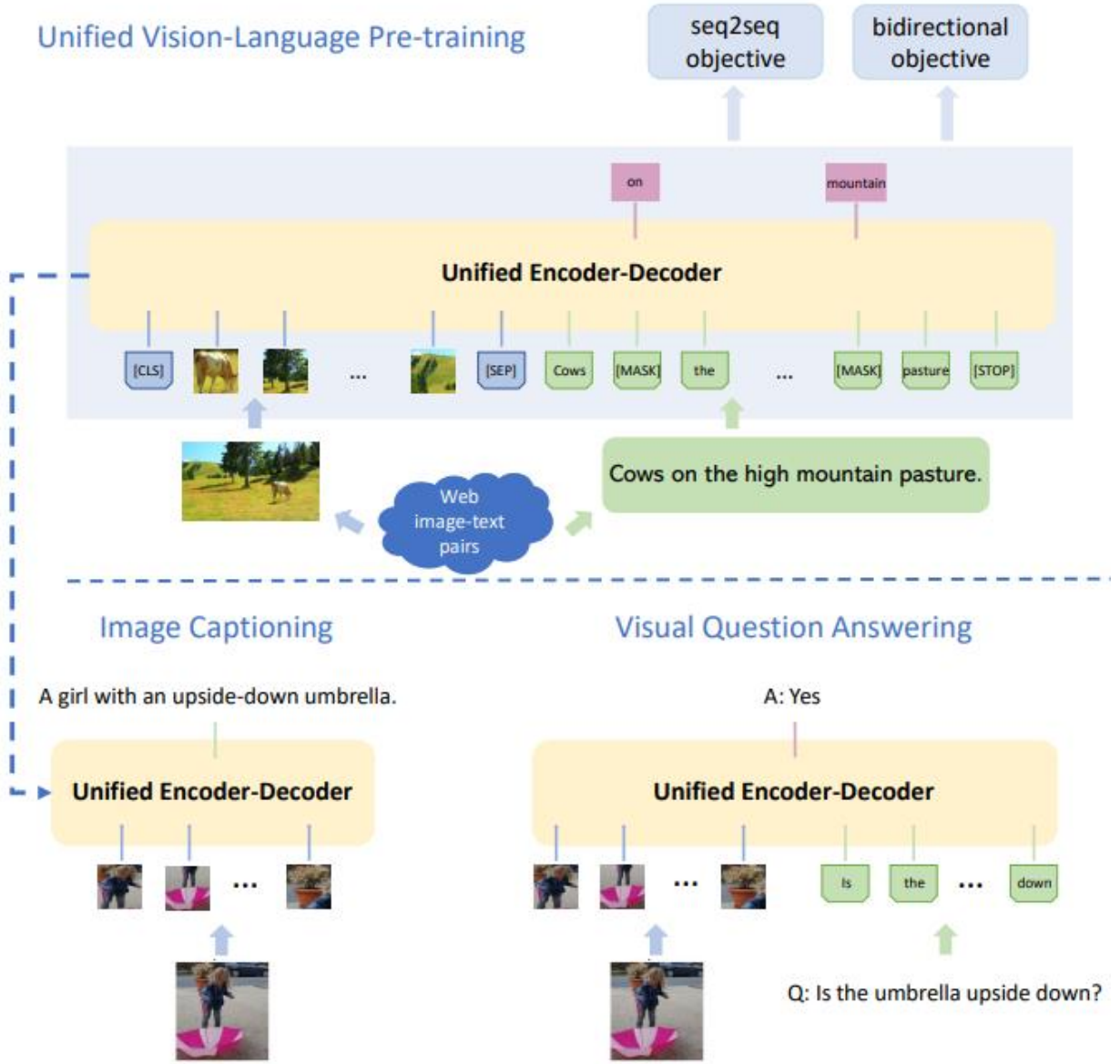
Vision-Language Transformer Model (pre-training)

- 15% of the text tokens are replaced by [MASK] token, random token or original token.
- The hidden state from the last transformer block is projected to word likelihoods where the masked token is predicted as classification problem.
- Through this reconstruction the model learns the dependencies in the context and forms a language model.
- Two objectives are considered within this model:
  - Bidirectional: every token can attend to every other token.
  - Seq-to-seq: tokens cannot attend to future tokens. It satisfies the auto-regressive property.
  - During pre-training they alternate between the two objectives.



VLP model for image-captioning

- For image captioning we fine-tune using seq2seq objective.
- During inference
  - Encode image region along with special tokens ([CLS] and [SEP] tokens).
  - We then start the generation by feeding in the [MASK] token and sampling a work from word likelihood output.
  - Replace the [MASK] token in the input sequence with sampled word and add new [MASK] token to the end of sequence.
  - Generation terminates when [STOP] token is chosen.



Training strategies

Cross-Entropy Loss

- Most used objective for image captioning.
- The aim is to minimize the negative log-likelihood of the current word given the previous ground-truth words.
- The loss works at word level and optimizes the probability of each word without considering long range dependencies.

$$L_{XE}(\theta) = -\sum_{i=1}^{n} \log\left(P\left(y_i \mid y_{1:i-1}, \boldsymbol{X}\right)\right)$$

Training strategies

Masked Language Model

- Idea is to randomly mask a subset of input tokens and train the model to predict these masked tokens based on remaining tokens, both previous and subsequent.
- The model relies more on context making it more robust.
- Training on these models is much slower since they only train in masked tokens and not entire sentence.

Data

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
COCO [128]	Generic	132K	5	27K (10K)	10.5
Flickr30K [129]	Generic	31K	5	18K (7K)	12.4
Flickr8K [19]	Generic	8K	5	8K (3K)	10.9
CC3M [130]	Generic	3.3M	1	48K (25K)	10.3
CC12M [131]	Generic	12.4M	1	523K (163K)	20.0
SBU Captions [4]	Generic	1M	1	238K (46K)	12.1
VizWiz [132]	Assistive	70K	5	20K (8K)	13.0
CUB-200 [133]	Birds	12K	10	6K (2K)	15.2
Oxford-102 [133]	Flowers	8K	10	5K (2K)	14.1
Fashion Cap. [134]	Fashion	130K	1	17K (16K)	21.0
BreakingNews [135]	News	115K	1	85K (10K)	28.1
GoodNews [136]	News	466K	1	192K (54K)	18.2
TextCaps [137]	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives [138]	Generic	849K	1/5	16K (7K)	41.8

Evaluation

- Evaluating quality of generated text is a tricky and subjective task.
- Image captions are further complicated since the caption cannot only be grammatical and fluent but also needs to properly refer to the image.
- The best way to evaluate this is still human evaluations
  - Costly
  - Not reproducible
- Automatic methods compare generated captions against human-produced references and are usually defined for other NLP tasks.

Evaluation

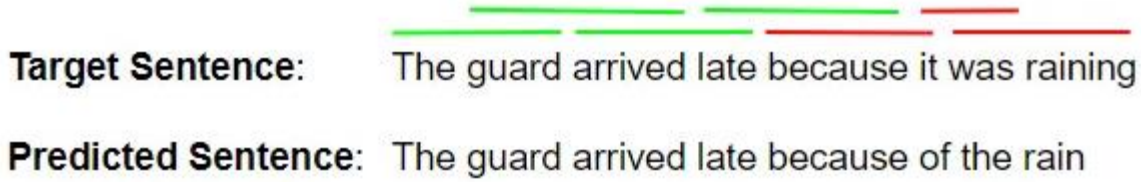
BLEU score

- Target sentence: The guard arrived late because it was raining
- Predicted sentence: The guard arrived late because of the rain

We first calculate precision scores for 1-gram through 4-grams.

$P_1 = 5/8$

$P_3 = 3/6$



$P_2 = 4/7$

$P_4 = 2/5$

Evaluation

BLEU score

- Brevity penalty: it penalizes sentences that are too short.

If the predicted sentence is just « the », the 1-gram precision is 1/1=1, indicating perfect score.

$$Brevity\ Penalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}$$

c = predicted sentence lenght

r = target sentence lenght

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

$$Geometric\ Average\ Precision\ (N) = (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}$$

Evaluation

ROUGE

- Compared to BLEU score that focuses on precision, ROUGE focuses on recall.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

TRAITEMENT AUTOMATIQUE DES LANGUES AVANCE

Master Informatique

2ème Année - 1er Semestre

Intervenants CM:

Gaël DIAS (gael.dias@unicaen.fr), Marc SPANIOL, Fabrice MAUREL

Navneet AGARWAL, Kirill MILINTSEVICH

GREYC Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

cnrs