

AMERICAN UNIVERSITY OF BEIRUT

AUTOMATING HUMAN'S COGNITIVE
PSYCHOLOGY FOR OPINION MINING
MODELS

by

OBEIDA AMER ELJUNDI

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
February 2021

AMERICAN UNIVERSITY OF BEIRUT

AUTOMATING HUMAN'S COGNITIVE
PSYCHOLOGY FOR OPINION MINING
MODELS

by
OBEIDA AMER ELJUNDI

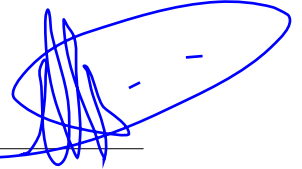
Approved by:



Dr. Hazem Hajj, Associate Professor

Advisor

Electrical and Computer Engineering, American University of Beirut



Dr. Imad Elhadj, Professor

Member of Committee

Electrical and Computer Engineering, American University of Beirut



Dr. Nizar Habash, Professor

Member of Committee

Computer Science, New York University Abu Dhabi

Date of thesis defense: January 27, 2021

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT
RELEASE FORM

Student Name: ElJundi Obeida Amer
Last First Middle

Master’s Thesis Master’s Project Doctoral Dissertation

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One ___ year from the date of submission of my thesis, dissertation or project.**
Two ___ years from the date of submission of my thesis , dissertation or project.
Three years from the date of submission of my thesis , dissertation or project.



Signature

February 8, 2021

Date

This form is signed when submitting the thesis, dissertation, or project to the University Libraries

Acknowledgements

This work wouldn't be possible without my advisor and my thesis committee. Special thanks for my committee, professor Imad Elhajj and professor Nizar Habash, for their useful feedback during my thesis proposal and for suggesting to limit the scope and focus on important and relevant tasks. I am also extremely thankful to my family, specially my parents, and my friends for their constant encouragement during my research period. Special thanks also to my colleague, Wissam Antoun, for his help with BERT implementation and reviewing the report. Last but not least, I owe a deep sense of gratitude to my advisor, professor Hazem Hajj, for all his technical and emotional support throughout my journey. Your dynamism, vision, sincerity, and motivation have deeply inspired me. I learned a lot from you professor, not only on the academic level, but also on the personal level. You are my role model I look up to. THANK YOU!

An Abstract of the Thesis of

Obeida Amer ElJundi for Master of Engineering
Major: Electrical and Computer Engineering

Title: Automating Human's Cognitive Psychology for Opinion Mining Models

This thesis focuses on the evaluation of automated reading comprehension of sentiment in text, which is also considered an opinion mining classification task. Previous work for opinion mining uses feature engineering machine learning (ML) or deep learning (DL) without consideration for the method's adaptation to the human's cognitive process. The aim of this thesis is to determine whether machines can learn better by following the human cognitive reading process or rather follow a machine-specific representation. The main difference lies in the intermediate representation of the data before classification.

On the human side, and based on recent psychological studies, it has been determined that reading comprehension is not the result of one single process as was thought before 1970s. Instead, psychologists realized that a combination of several complex cognitive processes are involved. Based on Cognitive Psychology, comprehension heavily depends on inference and background knowledge to construct a Situation model, which is a mental representation of the text. In other words, humans develop an image of the context in their minds before concluding its meaning. To emulate the Human Mental Intermediate (HMI) representation, we propose a Text-to-Image-to-Task (T2I2T) model comprehension by first mapping the input text to an image which provides a semantic representation equivalent to the mental representation of the text being analyzed.

From the machine's learning perspective, we conjecture that machines may not need to learn human-specific representations. Instead, we propose to explore the machine's ability in developing its own Machine Intermediate Internal (MII) representations through direct end-to-end (E2E) models. To emulate the machine's cognitive process, MII, we propose the use of Transformer-based Language Models (LM) and show that pre-training acts as a suitable means for the machine to acquire background knowledge comparable to the cognitive psychological Situation model.

To compare the performances of HMI and MII E2E model, we conducted experiments with applications to sentiment analysis. We developed our own data set with text-image-sentiment annotations by augmenting an exiting image captioning dataset with automated sentiment annotations. Several base models were developed for comparison including Bidirectional LSTMs with word embeddings and state of the art pre-trained LMs, such as BERT and ULMFit. The results showed that the machine’s E2E cognitive approach, MII, outperforms both LSTMs with word embeddings models and the human’s cognitive T2I2T approach, HMI, by 6% and 26% respectively.

The thesis also explored models to represent HMI and MII for Arabic. In particular, we developed models for HMI Image Captioning in Arabic and an Arabic MII universal language model, called hULMunA. In Arabic Image Captioning (AIC), we developed the first Arabic dataset and encoder-decoder end-to-end models to show that it is necessary to build language specific datasets and end-to-end models rather than translating English captions. In hULMunA, we developed the first Arabic specific Language Model and fine-tune it to achieve state-of-the-art results on four Arabic Sentiment Analysis datasets.

Contents

Acknowledgements	v
Abstract	vi
1 Introduction	1
2 Literature Review	4
2.1 Cognitive Models for Reading Comprehension	4
2.2 Computational Models for Reading Comprehension	5
3 Background: Discourse Comprehension in Cognitive Psychology	7
4 Machine Intermediate Representation for Situational Model with Application to Sentiment Analysis	10
4.1 Human Mental Intermediate (HMI) representation	11
4.1.1 HMI: Image to Sentiment	11
4.1.2 HMI: Image to Labels to Sentiment	11
4.1.3 HMI: Image Captioning Middle Representation	13
4.2 Machine Intermediate Internal (MII) Representation	14
5 Evaluating Machine Intermediate Representation: HMI versus MII for Sentiment Analysis	16
5.1 Dataset Preparation	16
5.2 Baseline	17
5.3 Results	18
6 Applications and Evaluation of HMI with Arabic Image Captioning	20
7 Applications and Evaluation of MII with Arabic End-to-End Language Model	24
8 Conclusion	28

List of Figures

3.1	Models of the reading process. [1]	8
3.2	Mental representation levels of discourse comprehension. [2]	9
3.3	The Situation model helps readers to infer additional information that is not explicitly stated in the text, such as the frog shape and its long tongue grabbing the bug.	9
4.1	Visual Sentiment Analysis by fine-tuning a ResNet pre-trained on ImageNet.	12
4.2	Detecting labels from an image using Google Vision API.	13
4.3	Our approach of automatically extracting text labels from an image (using Google Vision API), then mapping labels to sentiment.	13
4.4	A general overview of an encoder-decoder Image Captioning architecture.	14
4.5	Visual Sentiment Analysis utilizing Image Captioning middle representation using OSCAR.	15
5.1	Examples of our sentiment augmented COCO dataset.	18
5.2	Interpretability of our VSA system. Model attended to details insignificant for sentiment (e.g., background). This explains the poor performance for HMI.	19
6.1	AIC against translated EIC.	20
6.2	Sequence-to-Sequence Encoder-Decoder framework for Arabic Image Captioning.	21
6.3	BLEU scores of end-to-end AIC vs translating English captions.	23
6.4	Accurate results generated by our AIC model.	23
7.1	Three-step Process for Creating hULMonA.	26

List of Tables

5.1	Comparison of results (Accuracy F1) of baseline, HMI, and MII.	18
7.1	preprocessing reduces lexical sparsity	26
7.2	generating text using the pre-trained Arabic language model. . . .	27
7.3	Datasets statistics	27
7.4	Comparison of results (F1 Accuracy) obtained using hULMonA and other state-of-the-art models.	27

Chapter 1

Introduction

Reading comprehension is one of the most essential skills we need in our daily life. Although there have been tremendous achievements in the field of NLP in the last decade, most work did not account for the human reading (cognitive) processes. We study Cognitive Psychology, a sub-field of psychology concerned with human mental process, to grasp the fundamentals of reading comprehension. Cognitive psychologists believe that the output of reading comprehension is a mental representation of the text being read. The resultant mental representation is not the product of a single cognitive process. Rather, it involves many separate cognitive mechanisms [3] [4]. As a result of these various processes, three levels of representation are constructed during reading [3]. Introduced by Kintsch and van Dijk [5], the construction of the third and final representation level, the Situation model, depends heavily on inference and reader's background knowledge to go beyond what is explicitly stated in the text. Most recent reading comprehension theories can be traced to this seminal work of Kintsch and van Dijk.

Artificial Intelligence (AI) aims at mimicking human abilities to achieve tasks that requires human intelligence. Convolutional Neural Networks (CNN), for example, is inspired by some of the early findings about the human's visual system, particularly, the receptive fields in our visual cortex [6]. Motivated by our memory, Long Short-Term Memory (LSTMs) improved the performance of Recurrent Neural Networks (RNN) by introducing 1. memory gates to remember important and long dependencies and 2. and forget gates for unnecessary details [7]. Attention mechanisms mimic our ability of attending to the most salient parts of an image or text to achieve state-of-the-art results in both Computer Vision (CV) and Natural Language Processing (NLP). Several early psychological reading comprehension models [8, 9, 10, 11, 12, 13, 14, 15, 16] and recent NLP methods, including RNNs, cover only the first two cognitive representation levels but render the final and the most significant level, the Situation model, unaccounted for.

To mimic the human's cognitive Situation model, we propose two representation approaches. First, we develop the Human Mental Intermediate (HMI)

representation approach. HMI is inspired by the way the human mind develops an image that represents the context or situation of the text being read. The HMI model involves a Text-to-Image-to-Task (T2I2T) model by first mapping the input text to an image which provides an extended semantic representation equivalent to the mental Situation representation of the text being analyzed. In this thesis, we only focus on modeling the image to task. For the second approach, we propose to explore machine’s ability in developing its own Machine Intermediate Internal (MII) representations through direct end-to-end (E2E) models. Particularly, we utilize pre-trained Transformer-based Language Models (LM). We then compare both approaches to conclude the superior method of construing the best text representation that is equivalent to the cognitive Situation model. We evaluate the effectiveness of HMI strategy versus MII strategy with Opinion Mining as a case study.

Opinion Mining, also known as Sentiment Analysis, refers to the task of automatically extracting people’s opinions from digital text. Sentiment Analysis started gaining remarkable attention with the exponential growth of the online subjective data generated by users in form of text [17]. Sentiment analysis applications spread across multiple domains, including business and politics, providing insights into public opinion regarding policies, trends, or products [18].

In addition to the comparative case study, we explore the applications of HMI and MII models to Arabic. In particular we develop HMI Arabic models for Arabic captioning of images and MII Arabic universal Language model called hULMonA.

The thesis contributions can be listed as follows:

- First, we study reading comprehension in Cognitive Psychology to identify gaps in recent NLP methods.
- Second, we emulate the Human Mental Intermediate (HMI) representation by proposing mapping text to images to provide an extended semantic representation (Text-to-Image-to-Task (T2I2T)).
- Third, we explore machine’s ability to develop its own Machine Intermediate Internal (MII) representations through direct end-to-end (E2E) models.
- Forth, we develop a text-image-sentiment dataset to conduct a comparative analysis of HMI versus MII.
- Finally, we explore the applications of HMI and MII models to Arabic by developing Arabic Image Captioning (AIC) models and pre-training Arabic-specific Language Models.

This thesis is organized as follows. Chapter 2 reviews the literature of theoretical and computational models of reading comprehension in the field of Cognitive

Psychology. In chapter 3, we explain reading comprehension in Cognitive Psychology and lay the foundation of the situation model. We then propose our approaches of emulating the Situation model in chapter 4. We talk about our developed dataset and show our proposed methods results in chapter 5. Finally, we show some of our other applications related to the cognitive approach in chapter 6 and 7, and we conclude our thesis in chapter 8.

Chapter 2

Literature Review

Cognitive psychologists proposed many theories to explain the cognitive processes involved in reading comprehension. Moreover, computational implementations are developed to simulate some of the proposed theories. Both theoretical and computational models of discourse comprehension are reviewed here.

2.1 Cognitive Models for Reading Comprehension

There is a consensus in the cognitive psychology research community that the outcome of the reading comprehension processes is a mental representation of the text [19]. Nevertheless, the approach of constructing the resulting comprehension mental representation is still debatable. Over the last four decades, several reading comprehension models have been proposed by cognitive psychologists to conceptualize the construction process of the mental representation during and after reading. The cognitive reading process typically involves multiple steps of understanding, including: word level, sentence level, and complete document level.

Most proposed models describe only one aspect of the cognitive reading process. For word level, Interactive-Activation [8], Multiple-Levels [9], Dual Route Cascaded [10], and Bayesian Reader [11] models account only for word identification. For sentence level processing and syntactic parsing, Ferreira and Clifton [12] and Frazier [13] [14] proposed Garden-path model, and Jurafsky [15] and MacDonald et al [16] introduced the constraint-based model. At the document level, discourse processing models connects individual sentences into more global representation, named situation model, which reflects the overall comprehension of the text. Examples include the Construction-Integration (CI) model [20] [21] and the Event-Indexing model [22].

There has been attempts to cover multiple aspects of the reading comprehension process. For instance, models proposed by Just and Carpenter [23] and

Rayner and Pollatsek [24] accounted for both processing text at multiple levels (e.g., words and clauses) and eye movements during reading.

The first systematic analysis of reading comprehension was the top-down processing approach [25]. The most cited reading model proposed by Goodman [25] characterizes reading as a guessing game where readers develop various expectations about the text to be read then sample enough information from the text to either confirm or reject their expectations. To accomplish this sampling efficiently, the reader skips part of the text and directs the eyes to the most likely places in the text to find useful information. Another perspective of the reading process is the bottom-up approach [26] [27]. In this view, readers comprehend text hierarchically, starting from the perception of single phonemes to words, clauses, sentences and finally the whole piece of discourse. Unlike the top-down model, none of the text is skipped during reading. The interactive reading model [28] attempts to combine the valid insights of bottom-up and top-down models.

These aforementioned models did not address the gap of understanding the multifaceted nature of reading. For example, unconscious inference during reading must be taken into consideration. To address this gap, several psychologists [5] explained that the comprehension process involves not only a mental representation of the text itself, but also concepts that go beyond what is explicitly stated in the text. This reflection of the comprehension process is known as the situation model, and there has been several theories to describe it. The construction-Integration (CI) model proposed by Kintsch [20] [21] is considered to be the most complete and well-formulated model of text comprehension. During the construction phase, a dumb (automatic, bottom-up) process activates all related knowledge, including both relevant and irrelevant knowledge. In the following stage, called the integration phase, readers engage inference and background knowledge to prime (deactivate) irrelevant activated information. Another prominent situation model in the Event-Indexing model [22]. It states that as we perceive narratives, we segment text into events. Events can be split into five indexes, or dimensions, namely, time, space, protagonist, causality, and intentionality. Events that share at least one index are connected in the reader's brain. The more the shared indexes, the stronger the connection.

Despite the availability of many comprehension models, the literature still lacks a complete model that accounts for all of the different components of reading [1].

2.2 Computational Models for Reading Comprehension

Computational models of psychological text comprehension play significant role in understanding psychological complexities of text comprehension and facilitating

communication among researchers within and across research areas [29].

Miller and Kintsch [30] computational model, which is based on Kintsch and van Dijk [5] model, consisted of two components: a chunking program that identified propositions and a coherence program that was concerned with local coherence. Kintsch [20] [21] computationally modeled his CI theory as a connectionist network of nodes and links between them. Nodes indicate propositions, where links indicate activations that are built during the construction phase and updated during the integration phase. Several computational models were built on top of the CI model such that each makes different assumptions about one or more of the components or parameters of the computational processing model. The Capacity-Constraint Construction Integration (3CI) model [31] [32], for example, examined an alternative conception of working memory processes. Other CI variations, such as the Landscape model [33] [34], may alter the learning algorithm or the basis of establishing connections among nodes in the connectionist network.

Although most of the aforementioned models, including CI and Event-Indexing, are concerned with narrative text, which is objective in nature, Wang et al., [35] extended the Event indexing model to capture the subjective dynamics of social media text. Indexter [36] is another computational model built on top of the Event-Indexing model. Indexter is concerned about not only the simple story structure, but also how the experiencer receives the narrative.

To simulate concept activations in the memory, Latent Semantic Analysis (LSA) [37] [38] [39] is one of the most prominent algorithms used in several computational models, including Kintsch CI model and another implementation of the CI model [40]. Given a huge corpus, LSA provides semantic representations for any word in an unsupervised manner.

Chapter 3

Background: Discourse Comprehension in Cognitive Psychology

Reading comprehension is one of the most essential skills we need in our daily life. Till this day, the brain processes of reading comprehension are not completely unraveled. However, psychologists believe that the outcome of the reading comprehension processes is a mental representation of the text [19]. This mental representation is the result of many separate cognitive mechanisms, rather than a single cognitive process [3] [4]. As a result of these various processes, three levels of representation are constructed during reading [3]. Surface level is concerned with the exact meaning of particular words being read. While reading a sequence of words, we separately extract the literal meaning of every word by retrieving its mental representation from our vocabulary bank. The second representation level, Textbase, also known as propositional representation, connects the previous mental representations of separate words to construct idea units explicitly stated in the text. These two levels are enough to comprehend what is explicitly stated in the text. In fact, as figure 3.1 shows, several early theories were based only on these two levels to explain comprehension [8] [9] [10] [11] [12] [13] [14] [15] [16]. However, to completely comprehend a discourse, readers have to connect the Textbase representation of the currently read text to their background knowledge to process out of context ideas and resolve ambiguities [41]. Hence, Kintsch and van Dijk [5] introduced the Situation model that goes beyond what is explicitly stated in text. In fact, most discourse comprehension theories can be traced to this seminal work of Kintsch and van Dijk.

The construction of the Situation model depends heavily on inference and reader's background knowledge. There are three main types of inferences in the context of discourse comprehension: logical, bridging, and elaborative [42]. When we read the word "*widow*", we can immediately infer that the text is talking about a woman. This type of inference that depends only on the meaning of the word

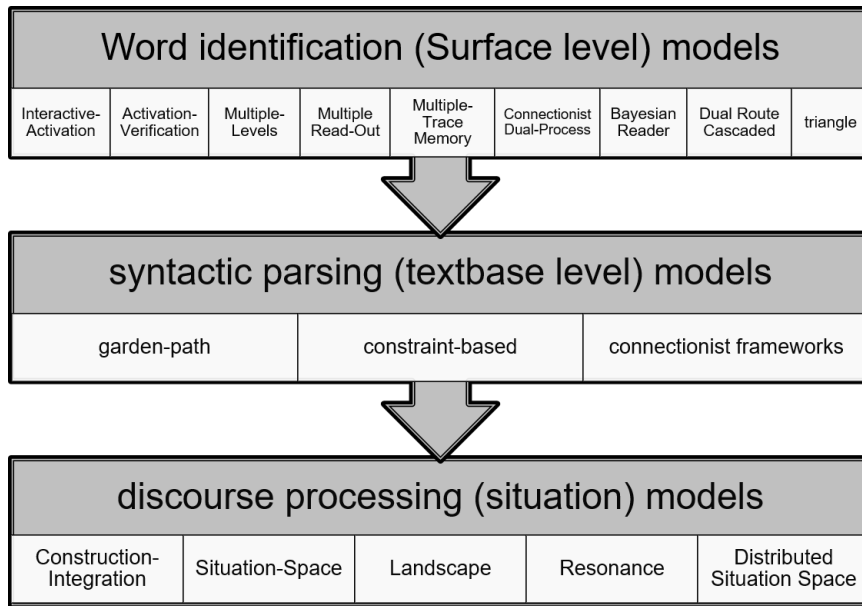


Figure 3.1: Models of the reading process. [1]

is called logical inference. Bridging inference, also known as backward inference, connects information in current environment (e.g., current sentence) with information previously stated. In elaborative inference, readers engage background knowledge to fill gaps in coherence and infer details not explicitly stated in text.

Consider this sentence as an example: *“the frog ate the bug”*. As figure 3.2 illustrates, after reading each word, we access the meaning from our semantic memory. This represents the Surface level representation. Then, we create relations between the words and create the first and only proposition in this example; EAT(FROG,BUG). This is known as the Textbase representation. Finally, the Situation representation helps us imaging the frog shape and its long tongue grabbing the bug; something similar to figure 3.3.

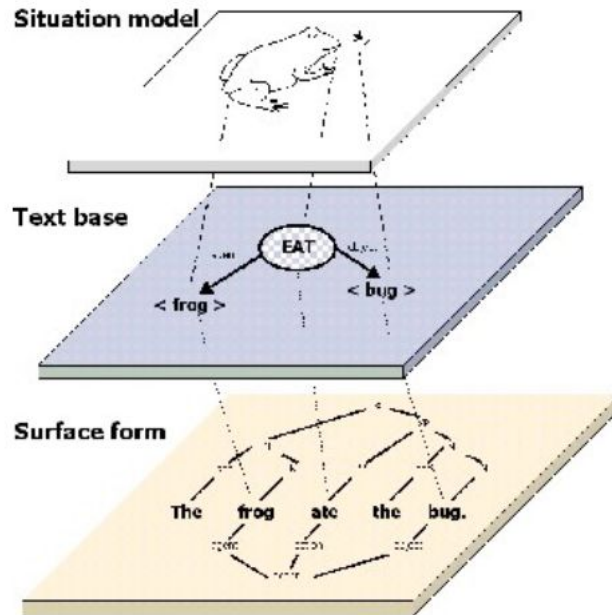


Figure 3.2: Mental representation levels of discourse comprehension. [2]



Figure 3.3: The Situation model helps readers to infer additional information that is not explicitly stated in the text, such as the frog shape and its long tongue grabbing the bug.

Chapter 4

Machine Intermediate Representation for Situational Model with Application to Sentiment Analysis

Any computational model for discourse comprehension must account for the three representation levels agreed on by the psychology community; the surface level representation, the textbase representation, and the situation model. Traditional deep learning approaches for text analysis involves Recurrent Neural Networks (RNN) [43] or one of its improved variations such as Long Short-term Memory (LSTM) [7]. Such models fall under the connectionist framework. As shown in figure 3.1, one limitation of such models is accounting only for surface level and textbase level processing. The most significant factor in discourse comprehension, the situation model, is often barely considered. Understanding cannot go beyond what is explicitly stated in the text for inference and background knowledge are taken for granted.

Situation model describes what the text is about. While reading, we tend to draw a picture in our head for the situation of the text being read. For instance, as shown in the example in figure 3.2, when we read *“the frog ate the bug”*, we end up imagining a frog trying to catch a flying bug with its long tongue. To account for the situation model, we propose to emulate the Human Mental Intermediate (HMI) representation by mapping text to an image that can be used to provide a new expanded semantic representation of the text. We then analyze the images using the state-of-the-art vision models (e.g., CNN) to infer sentiment. In section 4.1, we show three ways of inferring sentiment out of an image. We call this particular HMI approach of mapping text to image then to sentiment a Text-to-Image-to-Task (T2I2T).

Instead of forcing machines to use images as a middle representation, we also propose to explore machines ability in developing their own internal *“cognitive”*

representations through direct end to end (E2E) models. We call this Machine Intermediate Internal (MII) approach.

We'll first talk about our three ways of HMI, then we'll explicate MII.

4.1 Human Mental Intermediate (HMI) representation

In this section, we propose three different approaches of inferring sentiment out of an image.

4.1.1 HMI: Image to Sentiment

Studying the effect of images on humans, particularly concerning the evoked emotions, is a recent research area known as Visual Sentiment Analysis [44]. People recently tend to use images and visual content, besides the textual medium, on social media platforms to express their emotions; making Visual Sentiment Analysis an emerging field of study. Visual Sentiment Analysis can be formulated as image classification using deep learning methods such as Convolutional Neural Networks (CNN) [45, 46, 47]. CNN can process raw images as input to automatically extract relevant features for the purpose of classifying the sentiment expressed in the image (e.g., positive or negative). In fact, most of the state-of-the-art Visual Sentiment Analysis systems utilize transfer learning by fine-tuning a CNN pre-trained on a large, general dataset [48, 49, 50, 51].

Inspired by the previous work, we treat the task of classifying the sentiment expressed in our images as image classification task. Moreover, due to the relatively small size of our dataset, we utilize transfer learning to overcome overfitting. Namely, we fine-tune a pre-trained Residual network; one of the state-of-the-art models on ImageNet. ImageNet [52] is one of the largest and most widely used datasets in the field of Computer Vision. It contains more than 14 millions images with around 21 thousand groups or classes. We believe that pre-training a deep learning model on such dataset with large images and classes (e.g., objects) can be considered as a background knowledge for any other task, which is essential for constructing the situation model as described in chapter 3. Pre-trained on ImageNet, we fine-tune ResNet 101 [53]. ResNets utilize skip connections to prevent the problem of vanishing/exploding gradients in very deep neural networks. A high level overview of our Visual Sentiment Analysis system is shown in figure 4.1.

4.1.2 HMI: Image to Labels to Sentiment

Sentiment Analysis originally developed for the purpose of textual analysis [54]; hence, unlike Visual Sentiment Analysis, the field of textual Sentiment Analysis

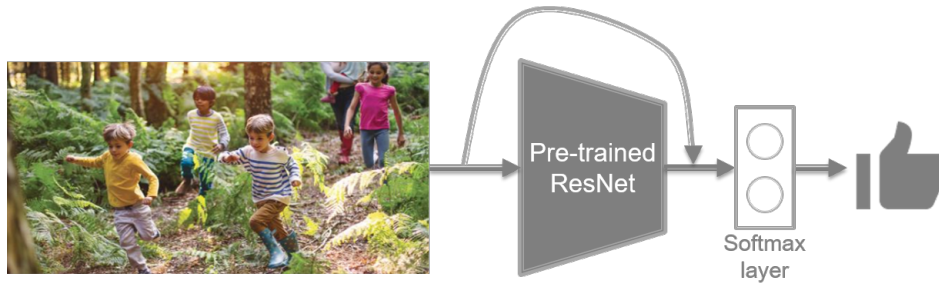


Figure 4.1: Visual Sentiment Analysis by fine-tuning a ResNet pre-trained on ImageNet.

has witnessed a significant improvement in the last decade [55] [56]. Therefore, we propose to map the image back to text, hoping to extract additional information from the image, and use a state-of-the-art system for textual Sentiment Analysis.

In the work of [57], authors claim that although some social media images are accompanied with text (e.g., title, description, and tags), this text cannot be exploited to extract the user’s sentiment. Authors demonstrated that the text accompanying an image is usually noisy and misleading as it contains camera related information (e.g., Nikon, D200), geographical information (e.g., Seattle), or objects that are not necessarily in the image. Moreover, the text might be subjective in a way that serves the user’s purposes, intentions, or agendas. Therefore, authors proposed to automatically extract objective text from an image using four different deep learning models. Two of the four models are object recognition models aiming at identifying objects in images.

Inspired by the previous work, we propose to map the image back to text to utilize the emerging progress of textual Sentiment Analysis systems. Using Google Vision API ¹, which is relying on ResNets, we automatically extract information about entities in an image, called labels, identifying general objects, locations, activities, animal species, products, and more. An example of labels automatically extracted from an image using Google Vision API is shown in figure 4.2. Detecting labels, such as activities, can provide information more than merely detecting object, which might be useful for sentiment analysis. For instance, the extracted *Play* and *Fun* activities in figure 4.2 can be directly linked to a positive sentiment. Such activity labels cannot be identified by an object detection system. We then fine-tune a pre-trained language model, namely ULM-FiT [58], using the extracted labels for the purpose of Sentiment Analysis. A high level overview of this approach is illustrated in figure 4.3.

¹<https://cloud.google.com/vision>

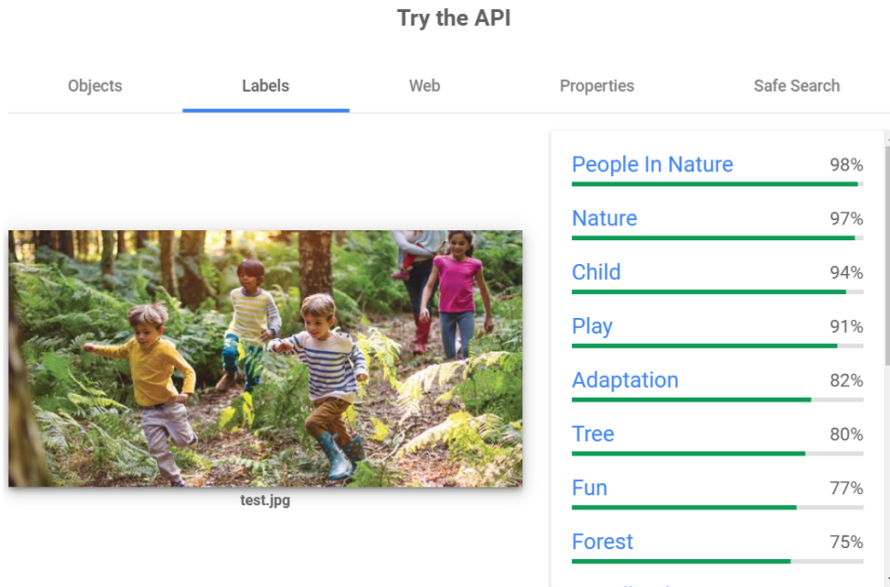


Figure 4.2: Detecting labels from an image using Google Vision API.

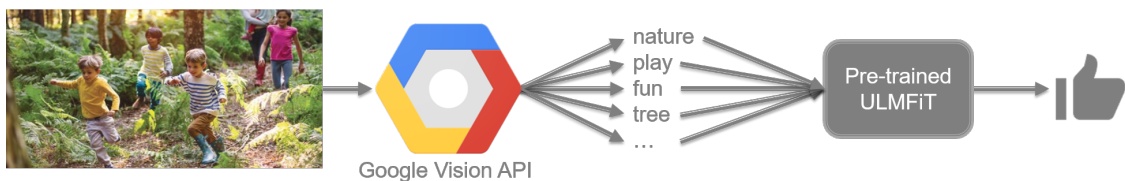


Figure 4.3: Our approach of automatically extracting text labels from an image (using Google Vision API), then mapping labels to sentiment.

4.1.3 HMI: Image Captioning Middle Representation

Identifying the relationships between the detected objects in an image can improve the overall performance of image understanding, and hence Sentiment Analysis. Image Captioning (IC) is the art of generating a human-readable sentence describing the content of an image. IC aims at not only detecting and recognizing objects in an image, but also understanding the interactions and relationships between the detected objects [59]. We hypothesize that this extended medium of understanding can improve the accuracy of any Computer Vision system, including the task we have in hand; Visual Sentiment Analysis.

The majority of the latest IC systems follows an encoder-decoder architecture. The encoder, usually a CNN, encodes the input image into a vector called image features or representation. Taking the image features vector as input, the decoder's objective is to generate a syntactically plausible and semantically meaningful sequence of words. Figure 4.4 demonstrates a general example of

an encoder-decoder IC system. In order for the decoder to successfully generate the image description, the encoder output, the image features vector, should embed all information about the detected objects and the relationships between them. We believe that the informative image features vector can be utilized for other tasks, such as Visual Sentiment Analysis, and can outperform extracting sentiment just from detected objects as described in section 4.1.2.

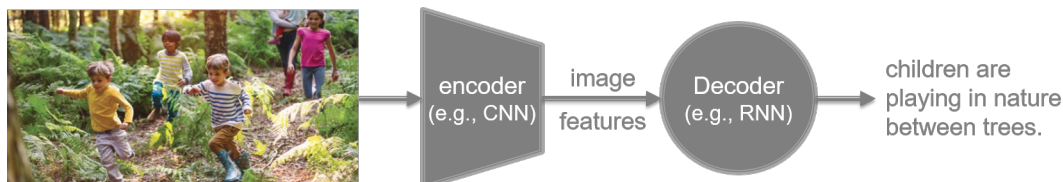


Figure 4.4: A general overview of an encoder-decoder Image Captioning architecture.

Learning cross-modal representations is essential for numerous Vision-Language (V+L) tasks, such as Image Captioning. Large-scale vision-language pre-training (VLP) using massive image-text pairs is becoming a popular trend to learn cross-modal representations for V+L tasks. Fine-tuning VLP models on downstream tasks achieved state-of-the-art results on well-established V+L tasks [60, 61, 62, 63, 64]. Object-Semantics Aligned Pre-training (OSCAR) [64] is one of the latest state-of-the-art methods for VLP. In addition to words sequence and image features, OSCAR leverages objects automatically detected in images, using Faster R-CNN [65], as anchor points to significantly ease the learning of image-text alignments; achieving state-of-the-art results on six well-established V+L tasks, including Image Captioning. To learn cross-modal *contextualized* representations, the input triple (word tokens, detected objects, and detected image regions) is fed to a multi-layer self-attention Transformer-based [66] encoder instead of a CNN. For the purpose of Visual Sentiment Analysis, as shown in figure 4.5, we will add a classification layer on top of OSCAR’s encoder, the pre-trained Transformer, to map the features vector to sentiment.

4.2 Machine Intermediate Internal (MII) Representation

In the previous section, we showed our approaches of adopting the situation physiological model for discourse comprehension by employing an image reflecting, and trying to go beyond, what is stated in the input text. We hope that an image will provide an extended knowledge environment machines can utilize to enhance natural language understanding. However, machines process images as rows and columns of pixels, a 3D matrix to be specific. One can argue that forcing machines to go through an image, that originally should be interpreted by the

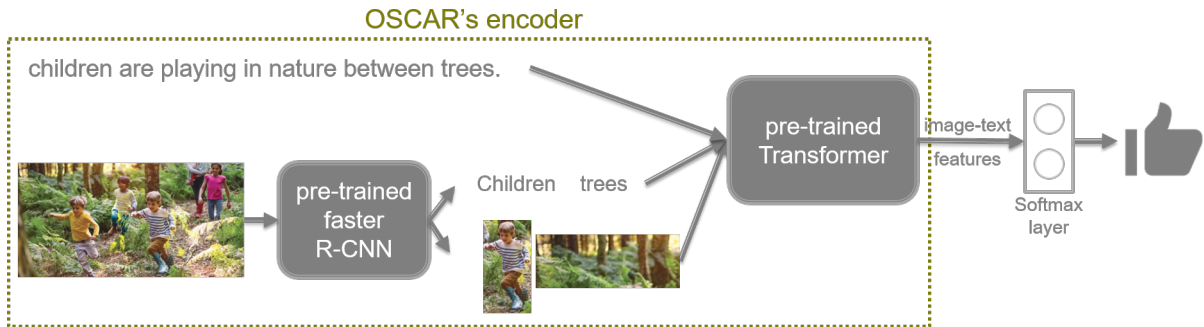


Figure 4.5: Visual Sentiment Analysis utilizing Image Captioning middle representation using OSCAR.

human visual system rather than machines, might limit machines performance and understanding.

For the last decade, it has been well-known that, using enough data, deep neural networks can automatically extract relevant features that outperforms the manually extracted features. Therefore, allowing machines to automatically extract its own middle representations directly from text should theoretically outperform the image middle representation approaches. This automatically extracted text middle representation can be thought of as the situation model for discourse comprehension in cognitive psychology, if the background knowledge is taken into consideration. As discussed in the introduction of chapter 4, the connectionist framework, including RNNs, fails to account for the situation model. One main reason is the lack of background knowledge and inference capabilities since RNNs were trained from scratch on limited data; only on a downstream dataset we have in hand.

To acquire background knowledge, we propose to use a model pre-trained on all available text (e.g., Wikipedia). We consider Language Modelling (LM) to be the ideal task to pre-train a model to obtain general understanding of a particular language due to its ability of capturing many aspects of language relevant for other downstream tasks, including sentiment orientation [67], hierarchical relations [68], and long-term dependencies [69]. We will use two of the state-of-the-art language model architectures for pre-training; namely ULMFiT [58], which uses AWD-LSTM [70], and BERT [71], which uses Transformer Encoders [66]. The pre-trained LM models can be fine-tuned on the task and dataset we have in hand by adding a classification layer on top of the pre-trained AWD-LSTM or Transformer Encoder.

Chapter 5

Evaluating Machine Intermediate Representation: HMI versus MII for Sentiment Analysis

In the chapter, we talk about how we built our own text-image-sentiment dataset, and why it was necessary to build a dataset from scratch. After that, we talk about a baseline model which our cognitive approaches, HMI and MII, will be compared against. Finally, we show and discuss our results.

5.1 Dataset Preparation

Our proposed human’s cognitive approach is consisted of two tasks: 1. mapping text to an image, and 2. inferring sentiment out of the image; also know as Visual Sentiment Analysis. The first task is achievable either by utilizing a dataset that already contains the text-image-sentiment triplet or by automatically generating an image by an end-to-end deep learning model.

Text-image-sentiment datasets are scarce and suffer from several limitations. In SentiCap [72], the authors developed a model that automatically generates image captions with positive or negative sentiments. They also built a text-image-sentiment dataset by assigning both positive and negative captions to every MS COCO images. Sentiment is hence reflected in the text only and is challenging to be extracted from images. Twitter for Sentiment Analysis (T4SA) [50] is another text-image-sentiment dataset. Authors collected around 3 million tweets containing both text and images and automatically predicted the sentiment polarity of the textual contents to train a visual sentiment classifier. We find that some Twitter text-image data, including T4SA, is impractical due to the following reasons. Text is sometimes not directly related or complementary to the accompanied image. Some images contain nudes and screenshots for games or chatting conversations, which is considered noisy for deep learning models. Fi-

nally, sentiment models developed for T4SA achieved low accuracy (51.3% on 3 classes) which reflects the poor quality of the dataset.

The second approach, on the other hand, is well-known in the literature as text-to-image synthesis, which aims at automatically generating realistic images from a text description. Most text-to-image synthesis relies heavily on Generative Adversarial Networks (GANs) [73]. Text-to-image synthesis is still one of the most challenging tasks in Computer vision. In fact most work in this field suffers from generalization and the limitation of generating images describing one particular domain only, such as flowers [74, 75], birds [74, 75, 76], bedrooms [77], etc. Therefore, relying on text-to-image synthesis to map texts to images is not feasible for our approach.

To test our T2I2T approach, it is necessary to build our own dataset of texts, their corresponding images, and sentiment. One simple way is to augment an Image Captioning dataset with sentiment as images accurately reflect what the text is talking about. We consider MS COCO [78]; a large scale, well-known, and reliable Image Captioning dataset. It contains around 330k images of complex everyday scenes containing common objects, 40k of which are considered for validation. Most images have more than one caption. Only the first caption is considered in our developed dataset. Sentiment was automatically assigned to the 40k validation captions using Google Natural Language API¹. The API processes the text and returns a sentiment score between -1 and 1 for every caption. Only sentiments with high confidence score are considered. To have a balanced dataset of 682 positive and 682 negative samples, we consider captions whose sentiment score is above 0.8 or below -0.69. Everything in between is treated as neutral sentiment and is not considered for our experiments. We believe that a total samples of 1364 are enough since we will be taking advantage of Transfer Learning in almost all our experiments. Examples of our sentiment COCO dataset are illustrated in figure 5.1.

5.2 Baseline

We compare our T2I2T and E2E approaches with Bidirectional LSTMs [79]; one of the most traditional and popular approaches before the emerge of the pre-trained LMs. We experiment with one layer of bidirectional LSTM with 256 neurons followed by a softmax layer of 2 neurons. Bidirectional LSTM inputs are represented as embeddings of length 300.



Figure 5.1: Examples of our sentiment augmented COCO dataset.

Proposed approach	Accuracy	F1
Baseline: LSTM	91.6	91.1
HMI: image→sentiment	71.5	72.3
HMI: image→labels→sentiment	56.5	56.8
HMI: IC representation	71.8	72.8
MII: ULMFiT	97.3	97.3
MII: BERT	97.4	97.4

Table 5.1: Comparison of results (Accuracy | F1) of baseline, HMI, and MII.

5.3 Results

Table 5.1 shows the results in terms of accuracy and F1 score of the LSTM baseline, our three HMI approaches, and two MII methods.

HMI. Using text labels (objects, activities, etc) as a middle representation deteriorates the performance of our model. In fact, going through text labels resulted in the lowest score. This suggests that converting an image to text leads to information and semantic loss. Going from image directly to sentiment achieved similar results as using IC middle representation. Apparently, both of their final representation looks the same, which means just like Image Captioning, image classification has the ability to extract not only objects in an image, but also the relationships between them. Nevertheless, 71.8% accuracy score is considered low on two classes classification task, and HMI performance is poor compared to the non-cognitive baseline (91.6% accuracy). This suggests that forcing machine to use images as a middle representation is not an efficient cognitive approach.

¹<https://cloud.google.com/natural-language>

One possible explanation is that images represent an informative environment for humans only since images are meant to be interpreted by our visual system. In brief, for humans, “*a picture is worth a thousand words,*” [80], but for a machine, an image is nothing but a 3D matrix of numbers. Another explanation can be concluded by studying the model’s interpretability. Figure 5.2 shows the model’s most salient regions of an incorrectly classified image. Instead of focusing on the two persons who most probably seem poor, sad, and depressed, the model attended to insignificant details such as the furniture and the background.



Figure 5.2: Interpretability of our VSA system. Model attended to details insignificant for sentiment (e.g., background). This explains the poor performance for HMI.

MII. Scoring almost the same results, both MII approaches, ULMFiT and BERT outperformed the baseline and the best HMI approach by around 6% and 26% respectively. Allowing machines to automatically utilize their own cognitive representations achieved better results than forcing a particular human representation.

Chapter 6

Applications and Evaluation of HMI with Arabic Image Captioning

In this chapter, we talk about one of our other applications related to MRI in Arabic. Particularly, we talk about Image Captioning in Arabic, which relates to MRI and going from image to text.

In our paper titled *Resources and End-to-End Neural Network Models for Arabic Image Captioning* [81], we answer the following question: to generate image captions in different languages, is it necessary to develop language-specific end-to-end models, or is it sufficient to translate English generated captions to destination language? We developed the first Arabic Image Captioning (AIC) end-to-end system. To evaluate its performance, we compared AIC with translating the captions of an English Image Captioning (EIC) system as shown in figure 6.1.

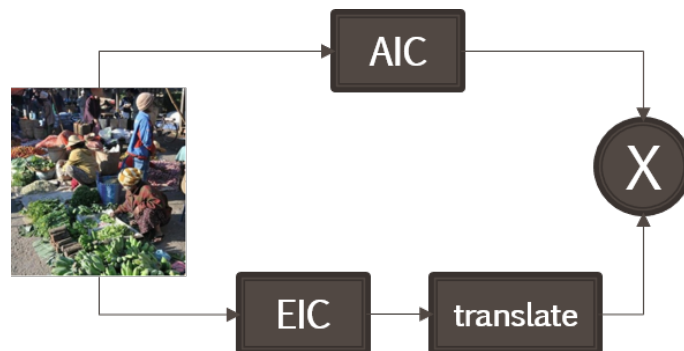


Figure 6.1: AIC against translated EIC.

As illustrated in figure 6.2, our AIC system follows the sequence-to-sequence encoder-decoder framework. For the encoder, we used CNN to encode the input

image to a fixed-length image vector. Instead of initializing our encoder CNN weights randomly and train from scratch, we will use the weights of a pre-trained CNN. This is known as transfer learning, which refers to the situation where what has been learned in one setting (task) is exploited to learn another setting (task). Transfer learning is used a lot in the literature to improve model generalization and speed up training. For our CNN, we use VGG16 [82], one of the previous state-of-the-art models for object detection. VGG16 contains thirteen convolution layers and three fully connected layers, and is able to detect approximately one thousand different objects.

For the decoder, we use RNN to decode the image vector into Arabic tokens. Particularly, we utilize LSTMs to overcome RNNs main issue of gradient vanishing during training due to its inability to handle long-term dependencies. LSTM inputs at different time stamps are represented by word embeddings, which are vectors of numbers that reflect semantics. Words with similar meaning have close embeddings. The embedding for each word is calculated as $x_t = W_e S_t$ for $t = 0, \dots, N$, where W_e is a $300 \times |V|$ word embedding matrix, meaning each word will be represented by a vector of length 300. $|V|$ denotes the vocabulary length, which is the number of unique words in our dataset. S_t is a $|V| \times 1$ one hot vector representing word i . Each hidden state of the LSTM emits a prediction for the next word in the sentence, denoted by $p_{t+1} LSTM(x_t)$.

Given any input image and its corresponding Arabic caption, the Arabic image captioning encoder-decoder model maximizes the following loss function: $\arg \max_{\theta} \sum_{(I,y)} \log p(y|I; \theta)$, where I is the input image, θ are parameters to be learned, and $y = y_1, \dots, y_t$ is the corresponding Arabic caption.

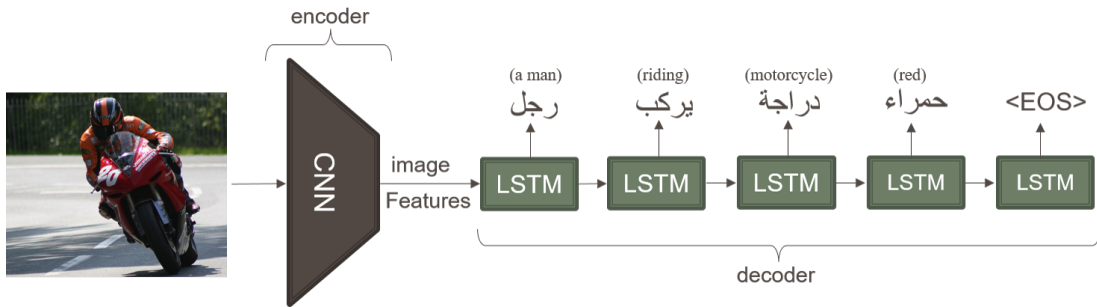


Figure 6.2: Sequence-to-Sequence Encoder-Decoder framework for Arabic Image Captioning.

To demonstrate the necessity of our end-to-end AIC system, we develop and train an English Image Captioning (EIC) system similar to our AIC using the original Flickr8K dataset. We then translate the generated English captions to Arabic using a pre-trained NMT model, namely Google Translate. The translated Arabic captions are evaluated and compared with our end-to-end AIC output. A

high level comparison of AIC against translated EIC is illustrated in Figure 6.1.

Due to the scarcity of AIC resources, we built, and made available for public, the first AIC dataset by translating a popular EIC dataset, namely Flickr8K [83], using Google Translate API¹. Flickr8K contains 8,000 images; each image has 5 captions and mainly showing humans and animals performing particular activity. Manual validation and editing for translated captions was done if necessary by professional Arabic translator to account for any incoherent translations.

Datasets contain raw text, which may include useless textual information. It is crucial to clean and preprocess our data before feeding it to any model because *'garbage in, garbage out'*. We followed Arabic preprocessing techniques recommended by [84]: Diacritics were removed, the *'hamza'* on characters was normalized, in addition to normalizing some word ending characters such as the *'t marbouta'* and *'ya' maqsoura'*. Moreover, we got rid of punctuation as well as non Arabic letters. Finally, a special start and end token were added at the beginning and the end of each caption to mark the starting and the ending point of each caption. Short captions were padded with a special padding token to ensure having captions of the same length.

For the image model, a pre-trained VGG16, excluding the last layer, was used to map images to embeddings, a vector of length 4096. The image embeddings vector was then mapped to a vector of 256 by a fully connected layer with tanh activation function to force the output values to be between -1 and 1. For the language model, a single hidden LSTM layer with 256 memory units was defined. The initial state of the LSTM was set to be the image embeddings, in order to ensure generating captions related to a specific image. The loss function was Softmax Cross Entropy. The optimization was done with mini batch Gradient Descent with Adam optimizer and batch size of 1024. The total number of epochs was 5. We consider an epoch as a single pass of the complete training dataset through the training process. Each epoch took around 25 seconds.

Following previous works, the model was evaluated on the BLEU-1,2,3,4 [85], which assesses a candidate sentence by measuring the fraction of n-grams that appear in a set of references. BLEU scores for our E2E AIC system versus translating EIC results are illustrated in figure 6.3. An end-to-end approach of directly generating Arabic captions outperformed translating English generated captions. One possible explanation is that using a deep learning model for English captioning followed by a second deep learning model for English-to-Arabic translation may accumulate both models errors and uncertainties. Figure 6.4 shows some examples of captions generated by our end-to-end AIC system. Dataset and code are available for public: <https://github.com/aub-mind/Arabic-Image-Captioning>

¹<https://cloud.google.com/translate>

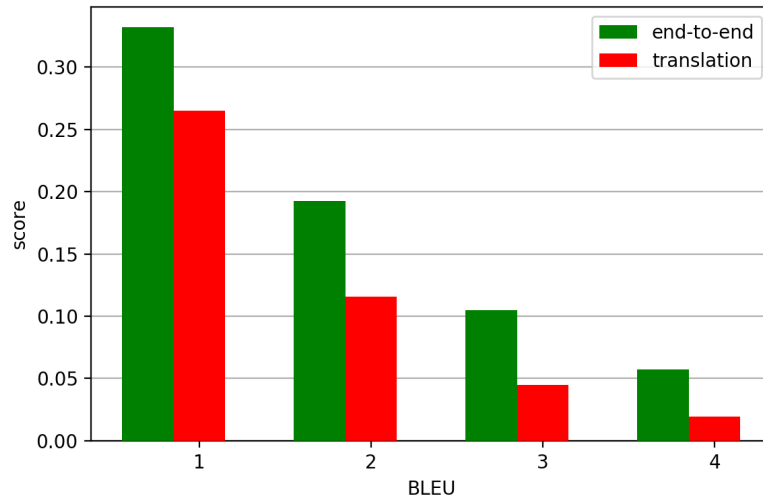


Figure 6.3: BLEU scores of end-to-end AIC vs translating English captions.



Figure 6.4: Accurate results generated by our AIC model.

Chapter 7

Applications and Evaluation of MII with Arabic End-to-End Language Model

In this chapter, we talk about one of our other applications related to MII in Arabic. Particularly, we show our work of pre-training the first Arabic-specific Language Model, which strongly correlates with MII.

Transfer Learning using Universal Language Models (ULMs), such as ULM-FiT [58] and BERT [71], have achieved state-of-the-art results in various NLP tasks in English. In the field of Arabic NLP, the use of Transfer Learning in Arabic has been mainly focused on word embedding models [86, 87]. In our work titled *hULMonA (حلمنا): The Universal Language Model in Arabic* [88], we hypothesize and prove that similar success can be achieved for Arabic. We developed the first Universal Language Model in Arabic (hULMonA - حلمنا meaning our dream), demonstrating its use for any Arabic classifications task.

Among the recently developed ULM models, BERT [71] built a multilingual language version using 104 languages including Arabic. One advantage of the multi-lingual BERT (mBERT) is that it can be used for many languages. However, one important limitation is that it was constrained to parallel multi-lingual corpora and did not take advantage of much larger corpora set available for Arabic, making its intrinsic representation limited for Arabic. As a result, there is an opportunity to further improve the potential for ULM success by developing

an Arabic specific ULM.

As figure 7.1 shows, hULMonA requires three steps. First, we pre-trained one of the state-of-the-art Language Models, namely AWD-LSTM, on the whole Arabic Wikipedia text (600K articles) to acquire general knowledge about the Arabic language. Although Wikipedia text is mainly in MSA, the resultant pre-trained model can be fine-tuned later on different text genres (e.g., tweets) and Arabic dialects to outperform training from scratch. Due to the huge amount of text and model parameters, especially at the last softmax layer which has as many neurons as the vocabulary size, the pre-training stage consumes much time and computational power. Fortunately, pre-training is done once, and the resultant model is made available to the community.

Second, to adapt to the new textual properties of the new (target) dataset, we fine-tuned our LM on the target dataset. This is crucial because although the general-domain LM is trained on MSA, most Arabic datasets and social media platforms contains dialects. Unlike MSA, dialects have no standard or codified form and are influenced by region specific slang. During fine-tuning, we use different learning rates for different layers, which is referred to as discriminative fine-tuning. This is crucial since different layers capture different types of information [89]. Discriminative fine-tuning updates the model parameters as follows:

$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta)$$

where θ^l is the model parameters of layer l , and η^l is the learning rate of layer l .

Finally, two fully connected layers are added to the LM for classification with ReLU and Softmax activations respectively. At first, the two fully connected layers are trained from scratch, while previous layers are frozen. After each epoch, the next lower frozen layer is unfrozen and fine-tuned until convergence. This is known as gradual unfreezing, and it is essential to avoid catastrophic forgetting of the information captured during language modeling.

hULMonA was constructed by first extracting and preprocessing all Arabic Wikipedia articles up to March of 2019. Articles images, links, and HTML were removed using an online tool¹, and articles with less than 100 characters were excluded resulting in 600,559 Arabic articles consisting of 108M words, 4M of which were unique. The large number of unique words requires more parameters to be learnt and is more prone to overfitting. This problem is called lexical sparsity, and it is a well-known challenge in Arabic NLP. Therefore, text was preprocessed by replacing numbers by a special token, normalizing Alif and Tamarbota, separating punctuations from words by a white space, and removing diacritics and non-Arabic tokens. Moreover, MADAMIRA [90], an Arabic morphological analyzer and disambiguator, was utilized to separate words prefixes, such as Al-taareef (the), and suffixes, such as possessive pronouns, resulting in words stems, thus, reducing lexical sparsity. Table 7.1 shows the number of

¹<https://github.com/attardi/wikiextractor>

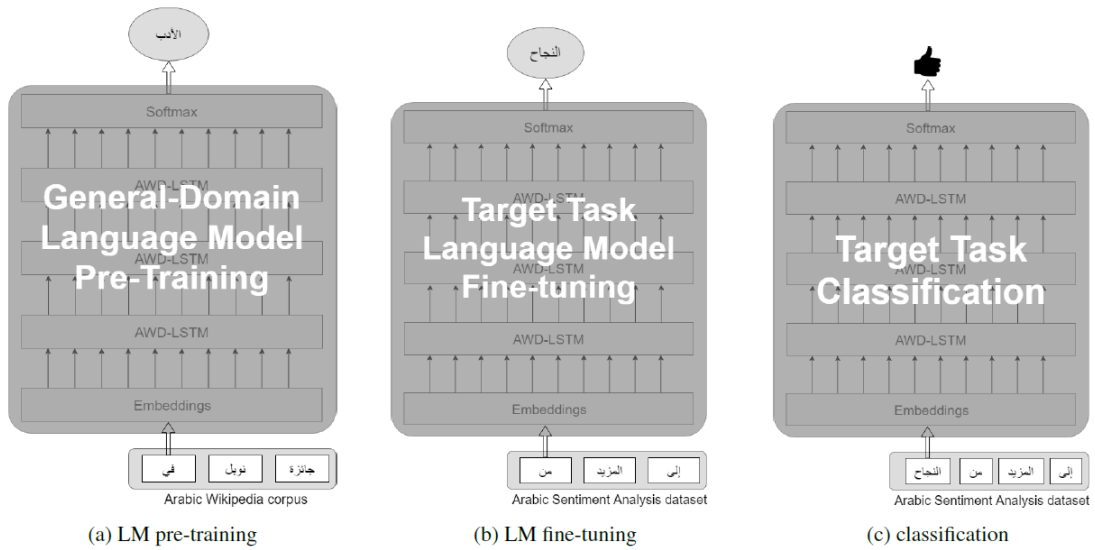


Figure 7.1: Three-step Process for Creating hULMonA.

unique words before and after preprocessing Arabic text using MADAMIRA. Finally, tokens that appeared less than 5 times were replaced by a special token.

	Example	Unique tokens
Before	الماء مادةٌ شفافةٌ عديمة اللون والرائحة	4.1M
After	ال+ ماء ماده شفاهه عديمه ال+ لون و+ ال+ راءحه	9.1K

Table 7.1: preprocessing reduces lexical sparsity

Table 7.2 demonstrates the capabilities of the pre-trained language model of generating coherent Arabic sequences based on initial tokens. To provide credible evaluation for the performance of the two ULM’s, we catalog a benchmark dataset for Arabic which can also be used for future research benchmark evaluations. The data sets vary in size allowing us to demonstrate the ULM’s abilities to fine tune with little data and achieve high performance. The benchmark data set is summarized in table 7.3 along with statistics on its content. Fine-tuning hULMonA achieved a new state-of-the-art on four Arabic Sentiment Analysis datasets as shown in table 7.4.

Initial tokens	Generated sequence
الدكتور (Doctor)	الدكتور احمد الحسن ، كاتب وباحث سعودي ، ولد في يونيو (Doctor Ahmad Al Hassan is a Saudi writer and researcher. He was born in June)
لاعب كرة قدم (football player)	لاعب كرة قدم امريكي يلعب كلاعب وسط (American football player plays as midfield)
وتقع دوله (The country is located)	وتقع دوله الامارات العربيه المتحده في الشرق الاوسط (United Arab Emirates is located in the middle east)

Table 7.2: generating text using the pre-trained Arabic language model.

Dataset	Resource	# samples	# classes	MSA Dialect
HARD [91]	Hotel reviews (www.booking.com)	93,700	2	MSA & Gulf
ASTD [92]	Twitter	10,000	4	MSA & Egyptian
ASTD-B [92]	Twitter	1,600	2	MSA & Egyptian
ArSenTD-Lev [93]	Twitter	4,000	5	Levantine Dialect

Table 7.3: Datasets statistics

Dataset	SOTA Results	hULMonA	mBERT
HARD	93.1 93.2 [91]	95.7 95.7	95.7 95.7
ASTD	62.0 68.7 [92]	67.7 69.9	67.0 77.1
ASTD-B	82.5 82.4 [94]	85.8 86.5	80.0 80.1
ArSenTD-Lev	50.0 51.0 [93]	51.1 52.4	51.0 51.0

Table 7.4: Comparison of results (F1 | Accuracy) obtained using hULMonA and other state-of-the-art models.

In conclusion, we utilized Transfer Learning to develop the first Arabic universal language model, hULMonA, that can be fine-tuned for almost any Arabic text classification task. Language knowledge learnt unsupervisedly from general-domain dataset is transferred to target task to improve overall performance and generalization. We show that hULMonA outperforms several state-of-the-art Arabic sentiment analysis datasets. In addition, we evaluate another ULM, mBERT, and compare results. We make hULMonA available for the community <https://github.com/aub-mind/hULMonA>

Chapter 8

Conclusion

We study Cognitive Psychology to find out that reading comprehension involves several cognitive processes, some of which are not covered by recent NLP methods. We developed models to account for the uncovered human’s cognitive processes by mapping text to images that can provide additional semantic environment. We develop and compare three methods of inferring sentiment from the image intermediate representation. We found that enabling machines to extract their own cognitive representations through end-to-end models outperforms going through images as an intermediate representation. We also build our own text-image-sentiment dataset to evaluate and compare our proposed approaches. We conclude that although images represent a rich environment of information for humans, utilizing images as an intermediate representation misled machine’s attention to focus on unrelated details, and hence end-to-end cognitive models achieved the best performance. No matter how advanced AI technologies get, they must be inspired by the most intelligent creature in the universe; humans.

We finally show two other applications of the human’s and the machine’s intermediate representation for Arabic NLP, namely Arabic Image Captioning and Arabic End-to-End Language Model. In Arabic Image Captioning, we address the challenge of Image Captioning in Arabic including the lack of Arabic resources. We develop a new Arabic Image Captioning dataset and propose two separate models for evaluation: translated English Image Captioning, and 2. end-to-end model that directly transcribes Arabic text from images. The models are compared using our developed dataset, and the results show the superiority of our end-to-end Arabic Image Captioning system. In Arabic end-to-end Language Model, we advance the Transfer Learning progress in Arabic by developing the first Arabic-specific universal Language model. Pre-trained on a huge data, the general-domain Language Model can be fine-tuned on any Arabic text classification task and any Arabic dialect. We show the superiority of our Language Model by achieving state-of-the-art results on four Arabic Sentiment Analysis datasets.

For our future work, we intend to conduct further analysis on the final representation of both approaches, HMI and MII, and visualize them in a 3D space in

our attempt to understand and explain models' behaviours. We are also planning to apply our findings of the reading comprehension cognitive process on a large variety of NLP tasks other than Sentiment Analysis to evaluate the generalization of our approach. Other tasks might include emotion recognition, cyberbullying and hate speech detection, sarcasm detection, fake news detection, etc. Finally, instead of studying the psychology of reading comprehension in general, we will focus on the psychological cognitive processes of particular downstream tasks, such as the mental processes involved in Sentiment Analysis specifically.

Appendix A

Abbreviations

AI	Artificial Intelligence
AIC	Arabic Image Captioning
CI	Construction-Integration
CNN	Convolutional Neural Network
CV	Computer Vision
EIC	English Image Captioning
E2E	End-to-End
DL	Deep Learning
GAN	generative Adversarial Network
HMI	Human Mental Intermediate representation
hULMonA	the first Universal Language Model in Arabic
IC	Image Captioning
LM	Language Model
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MII	Machine Intermediate Internal representation
ML	Machine Learning
NLP	Natural Language Processing
OSCAR	Object-Semantics Aligned Pre-training
RNN	Recurrent Neural Network
T4SA	Twitter for Sentiment Analysis
T2I2T	Text-to-Image-to-Task
ULM	Universal Language Model
V+L	Vision-Language
VLP	Vision-Language Pre-training

Bibliography

- [1] K. Rayner and E. D. Reichle, “Models of the reading process,” *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 1, no. 6, pp. 787–799, 2010.
- [2] Wikibooks, “Cognitive psychology and cognitive neuroscience/print version — wikibooks, the free textbook project,” 2017. [Online; accessed 20-October-2018].
- [3] T. A. Van Dijk, W. Kintsch, and T. A. Van Dijk, “Strategies of discourse comprehension,” 1983.
- [4] C. R. Fletcher, “Levels of representation in memory for discourse.,” 1994.
- [5] W. Kintsch and T. A. Van Dijk, “Toward a model of text comprehension and production.,” *Psychological review*, vol. 85, no. 5, p. 363, 1978.
- [6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] J. L. McClelland and D. E. Rumelhart, “An interactive activation model of context effects in letter perception: I. an account of basic findings.,” *Psychological review*, vol. 88, no. 5, p. 375, 1981.
- [9] D. Norris, “A quantitative multiple-levels model of reading aloud.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, no. 6, p. 1212, 1994.
- [10] M. Coltheart, K. Rastle, C. Perry, R. Langdon, and J. Ziegler, “Drc: a dual route cascaded model of visual word recognition and reading aloud.,” *Psychological review*, vol. 108, no. 1, p. 204, 2001.
- [11] D. Norris, “The bayesian reader: Explaining word recognition as an optimal bayesian decision process.,” *Psychological review*, vol. 113, no. 2, p. 327, 2006.

- [12] F. Ferreira and C. Clifton Jr, “The independence of syntactic processing,” *Journal of memory and language*, vol. 25, no. 3, pp. 348–368, 1986.
- [13] L. Frazier, “Sentence processing: A tutorial review.,” 1987.
- [14] L. Frazier, “Parsing modifiers: Special purpose routines in the human sentence processing mechanism,” *Comprehension processes in reading*, pp. 303–330, 1990.
- [15] D. Jurafsky, “A probabilistic model of lexical and syntactic access and disambiguation,” *Cognitive science*, vol. 20, no. 2, pp. 137–194, 1996.
- [16] M. C. MacDonald, N. J. Pearlmutter, and M. S. Seidenberg, “The lexical nature of syntactic ambiguity resolution,” *Psychological review*, vol. 101, no. 4, p. 676, 1994.
- [17] H. Chen and D. Zimbra, “Ai and opinion mining,” *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 74–80, 2010.
- [18] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [19] P. Kendeou, P. Van Den Broek, A. Helder, and J. Karlsson, “A cognitive view of reading comprehension: Implications for reading difficulties,” *Learning disabilities research & practice*, vol. 29, no. 1, pp. 10–16, 2014.
- [20] W. Kintsch, “The role of knowledge in discourse comprehension: A construction-integration model,” *Psychological review*, vol. 95, no. 2, p. 163, 1988.
- [21] W. Kintsch and C. Walter Kintsch, *Comprehension: A paradigm for cognition*. Cambridge university press, 1998.
- [22] R. A. Zwaan, M. C. Langston, and A. C. Graesser, “The construction of situation models in narrative comprehension: An event-indexing model,” *Psychological science*, vol. 6, no. 5, pp. 292–297, 1995.
- [23] M. A. Just and P. A. Carpenter, “A theory of reading: From eye fixations to comprehension.,” *Psychological review*, vol. 87, no. 4, p. 329, 1980.
- [24] K. Rayner, A. Pollatsek, J. Ashby, and C. Clifton Jr, *Psychology of reading*. Psychology Press, 2012.
- [25] K. S. Goodman, “Reading: A psycholinguistic guessing game,” in *Making Sense of Learners Making Sense of Written Language*, pp. 115–124, Routledge, 2014.

- [26] K. E. Stanovich, “Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy,” *Journal of education*, vol. 189, no. 1-2, pp. 23–55, 2009.
- [27] A. Garnham, *Psycholinguistics: central topics*. Methuen, 1985.
- [28] D. E. Rumelhart, *Toward an interactive model of reading*. International Reading Association, 1994.
- [29] S. R. Goldman, R. Golden, and P. van den Broek, “Why are computational models of text comprehension useful,” *Higher-level language processes in the brain*, pp. 27–51, 2007.
- [30] J. R. Miller and W. Kintsch, “Readability and recall of short prose passages: A theoretical analysis,” *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 4, p. 335, 1980.
- [31] S. R. Goldman and S. Varma, “Capping the construction-integration model of discourse comprehension,” *Discourse comprehension: Essays in honor of Walter Kintsch*, pp. 337–358, 1995.
- [32] S. R. Goldman, S. Varma, and N. Cote, “Extending capacity-constrained construction integration: Toward “smarter” and flexible models of text comprehension,” *Models of understanding text*, pp. 73–113, 1996.
- [33] P. Van den Broek, K. Risdén, C. R. Fletcher, and R. Thurlow, “A “landscape” view of reading: Fluctuating patterns of activation and the construction of a stable memory representation,” *Models of understanding text*, pp. 165–187, 1996.
- [34] P. Van den Broek, M. Young, Y. Tzeng, T. Linderholm, *et al.*, “The landscape model of reading: Inferences and the online construction of a memory representation,” *The construction of mental representations during reading*, pp. 71–98, 1999.
- [35] Y. Wang, D. Alahakoon, and D. De Silva, “An extended cognitive situation model for capturing subjective dynamics of events from social media,” *Australasian Journal of Information Systems*, vol. 22, 2018.
- [36] R. E. Cardona-Rivera, B. A. Cassell, S. G. Ware, and R. M. Young, “Indexer: A computational model of the event-indexing situation model for characterizing narratives,” in *Proceedings of the 3rd Workshop on Computational Models of Narrative*, pp. 34–43, 2012.
- [37] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

- [38] T. K. Landauer and S. T. Dumais, “A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.,” *Psychological review*, vol. 104, no. 2, p. 211, 1997.
- [39] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
- [40] B. Lemaire, G. Denhière, C. Bellissens, and S. Jhean-Larose, “A computational model for simulating text comprehension,” *Behavior research methods*, vol. 38, no. 4, pp. 628–637, 2006.
- [41] A. C. Graesser, K. K. Millis, and R. A. Zwaan, “Discourse comprehension,” *Annual review of psychology*, vol. 48, no. 1, pp. 163–189, 1997.
- [42] M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student’s handbook*. Psychology press, 2013.
- [43] J. L. Elman, “Finding structure in time,” *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
- [44] A. Ortis, G. M. Farinella, and S. Battiato, “Survey on visual sentiment analysis,” *IET Image Processing*, vol. 14, no. 8, pp. 1440–1456, 2020.
- [45] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, “Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks,” *arXiv preprint arXiv:1410.8586*, 2014.
- [46] C. Xu, S. Cetintas, K. chih Lee, and L. Li, “Visual sentiment prediction with deep convolutional neural networks,” *ArXiv*, vol. abs/1411.5731, 2014.
- [47] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” *arXiv preprint arXiv:1509.06041*, 2015.
- [48] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, “Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction,” in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pp. 57–62, 2015.
- [49] J. Islam and Y. Zhang, “Visual sentiment analysis for social images using transfer learning approach,” in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pp. 124–130, IEEE, 2016.

- [50] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell’Orletta, F. Falchi, and M. Tesconi, “Cross-media learning for image sentiment analysis in the wild,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 308–317, 2017.
- [51] V. Campos, B. Jou, and X. Giro-i Nieto, “From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction,” *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.
- [52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.
- [53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [54] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. 2, p. 1–135, Jan. 2008.
- [55] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.
- [56] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, “Deep learning for aspect-based sentiment analysis: a comparative review,” *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019.
- [57] A. Ortis, G. M. Farinella, G. Torrioni, and S. Battiato, “Exploiting objective text description of images for visual sentiment analysis,” *Multimedia Tools and Applications*, pp. 1–24, 2020.
- [58] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” *arXiv preprint arXiv:1801.06146*, 2018.
- [59] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.
- [60] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.
- [61] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “Vl-bert: Pre-training of generic visual-linguistic representations,” *arXiv preprint arXiv:1908.08530*, 2019.

- [62] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, 2019.
- [63] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, “Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.,” in *AAAI*, pp. 11336–11344, 2020.
- [64] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *European Conference on Computer Vision*, pp. 121–137, Springer, 2020.
- [65] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [67] A. Radford, R. Jozefowicz, and I. Sutskever, “Learning to generate reviews and discovering sentiment,” *arXiv preprint arXiv:1704.01444*, 2017.
- [68] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni, “Colorless green recurrent networks dream hierarchically,” *arXiv preprint arXiv:1803.11138*, 2018.
- [69] T. Linzen, E. Dupoux, and Y. Goldberg, “Assessing the ability of lstms to learn syntax-sensitive dependencies,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 521–535, 2016.
- [70] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing lstm language models,” *arXiv preprint arXiv:1708.02182*, 2017.
- [71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [72] A. Mathews, L. Xie, and X. He, “Senticap: Generating image descriptions with sentiments,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.
- [73] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

- [74] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” *arXiv preprint arXiv:1605.05396*, 2016.
- [75] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.
- [76] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X.-Y. Jing, “Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis,” *arXiv preprint arXiv:2008.05865*, 2020.
- [77] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.
- [78] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [79] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
- [80] A. N. Hibbing and J. L. Rankin-Erickson, “A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers,” *The reading teacher*, vol. 56, no. 8, pp. 758–770, 2003.
- [81] O. ElJundi, M. Dhaybi, K. Mokadam, H. M. Hajj, and D. C. Asmar, “Resources and end-to-end neural network models for arabic image captioning,” in *VISIGRAPP (5: VISAPP)*, pp. 233–241, 2020.
- [82] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [83] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.
- [84] A. Shoukry and A. Rafea, “Preprocessing egyptian dialect tweets for sentiment mining,” in *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, p. 47, 2012.

- [85] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- [86] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, “Word embeddings and convolutional neural network for arabic sentiment classification,” in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 2418–2427, 2016.
- [87] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, “Aravec: A set of arabic word embedding models for use in arabic nlp,” *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.
- [88] O. ElJundi, W. Antoun, N. El Droubi, H. Hajj, W. El-Hajj, and K. Shaban, “hulmona: The universal language model in arabic,” in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 68–77, 2019.
- [89] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in neural information processing systems*, pp. 3320–3328, 2014.
- [90] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, “Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic,” in *LREC*, vol. 14, pp. 1094–1101, 2014.
- [91] A. Elnagar, Y. S. Khalifa, and A. Einea, “Hotel arabic-reviews dataset construction for sentiment analysis applications,” in *Intelligent Natural Language Processing: Trends and Applications*, pp. 35–52, Springer, 2018.
- [92] M. Nabil, M. Aly, and A. Atiya, “Astd: Arabic sentiment tweets dataset,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2515–2519, 2015.
- [93] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. Shaban, “Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (H. Al-Khalifa, K. S. University, K. W. Magdy, U. of Edinburgh, U. K. Darwish, Q. C. R. Institute, Q. T. Elsayed, Q. University, and Qatar, eds.), (Paris, France), European Language Resources Association (ELRA), may 2018.
- [94] A. Dahou, M. A. Elaziz, J. Zhou, and S. Xiong, “Arabic sentiment classification using convolutional neural network and differential evolution algorithm,” *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

