Università degli Studi di Milano
Facoltà di Scienze e Tecnologie
Corso di Laurea in Informatica (classe lm-18)

MULTIMODAL EMOTION ANALYSIS FOR PRAGMATIC INTERPRETATION OF WORDS IN CONTEXT

Relatore: Prof. Giuseppe BOCCIGNONE

Correlatore: Dr. Sathya BURSIC

Tesi di:
Alessandro BALLERINI
Matricola: 927412

Anno Accademico 2021-2022

Contents

Chapter 1

Introduction

When people communicate with each other, often, what is said and what is meant do not coincide; it is normal for a person to infer the meaning behind incomplete utterances, rhetorical figures, implicatures, and other non-direct ways of verbal communication. It is easy for a person to "read the room" or "read between lines" during a conversation or, in general, in the course of social interaction, but why is it so, and which mechanisms allow implicatures in non-explicit contexts? This thesis aims at building a system able to integrate different sources of information and derive the correct interpretation of an utterance in non-trivial circumstances, such as the case of an interaction that comprehends irony or other figures of speech. We will explore the role of emotions in pragmatic reasoning and how much and in which way they contribute to the communicative act.

Motivations Language alone is often not enough to convey meaning; a shared background knowledge is fundamental between the actors involved in the communicative act. The common background may be given for granted by the speakers in order to ease the interaction by removing information that may be considered redundant; it is nonetheless essential in the communicative act, and its absence or reduction may cause misunderstandings or the impossibility of communication. People may also use complex figures of speech such as irony, metaphor, or hyperbole to communicate non-trivial concepts or effectively maintain the listeners' attention. Additional sources of information (prosody, body language, facial expressions) are integrated to disambiguate or enhance aspects of the communication.

The expert use of the voice helps the speaker maintain the focus of a crowd on the subject; he may underline the crucial aspects by supporting them with a more imposing stance and voice tone and can reduce the intensity of his speech when dealing with minor topics. The actor on stage exploits prosody and body language to convey emotions to the audience and keep them involved in the development of the plot.

All those elements introduce us to the concept of pragmatic. Pragmatic concerns the insertion of language into context, reflecting the belief that language alone is not an efficient and effective vehicle for conveying unambiguous meaning; it allows us to distinguish between what is said and what is meant by going beyond the literal meaning of words.

In particular, emotions represent one of the primary vehicles of information; They allow the understanding of complex situations involving abstract forms of speech, such as irony, where the literal interpretation of a speaker's utterance may convey the opposite of the desired meaning, or hyperbole, where the speaker deliberately exaggerates some elements of the speech to the point of absurdity.

As an example of the importance of emotions in our daily communications, we often use emojis in text chats to disambiguate written sentences and help us be more concise by expressing our emotional state about the matter without making it explicit.

Main contributions The work presented here is grounds in the Rational Speech Act framework for pragmatic reasoning [1] and tries to expand the model by introducing emotions as a source of information to interpret ambiguous utterances more precisely. In addition, a generative deep probabilistic model is used to extract emotional clues from the speaker's facial expressions, to use in the simulation of the communicative act.

The goal is to show and quantify how the introduction of emotions in the communicative act allows the agents to convey meaning in a better and less ambiguous way. To reach our goal, we start by building an instance of the rational speech act to perceive the concept of irony in a dialogue between two actors. This first instance represents the starting point of our analysis to measure our improvements. Next, we will introduce the concept of emotions. To do so, we will use a generative model, a Variational Autoencoder (VAE)[2], to represent emotions in a twofold way: as a categorical entity such as anger, happiness, sadness, and other discrete concepts, and as an image of an individual expressing such emotion. The generative model allows us to "translate" one modality into the other, which means we can classify the emotion expressed by one person's facial expressions and generate an image of a person expressing the desired emotion. As we will see in the following chapters, we built different models considering different modalities for representing emotions. Our first approach consists of training the model using raw images of people's faces using complex convolutional neural networks to extract and encode the features of people's faces. The second approach was to extract action units from the raw images using an external tool called OpenFace [3] and then use those values as a modality in our variational autoencoder. Finally, we use another external tool called OpenFACS [4] to generate an image expressing the emotion represented by the model's action unit

values. We eventually introduce emotions into the Rational Speech Act by considering the problematic case of irony and measuring the improvements generated by our addition.

Organization of the work This thesis unfolds as follows.

- **Chapter One** covers a brief introduction to the subject and objectives of this thesis.
- **Chapter Two** will define the concept of emotion, the biological/psychological basis, and how it is represented in a computable form. We also briefly introduce the main theories of emotions, focusing on Basic Emotion Theory, constructivist theories, and Conceptual Act Theory.
- **Chapter Three** we introduce the theoretical bases of the models we implemented. We describe the Rational Speech Act framework, which utilizes recursive reasoning and an agent-based approach to make inferences about non-observed random variables representing the internal state and the intentions of two human agents involved in the communicative act. We focus on the mathematical definition of the Rational Speech act Framework, considering its general functioning and application for the modeling of irony. In doing so, we focus on how introducing emotion in the communicative act allows for clarify non-explicit interactions, improving the pragmatic interpretation of the spoken words.
- **Chapter Four** illustrates the implementations of the models described in chapter four. We also propose a way of extracting and modeling emotional concepts through a generative approach. We use a Multimodal Variational Autoencoder to build a latent representation of the emotion concept, which can represent the same emotional state in a discrete category or the form of the corresponding facial expression. We finally use the MVAE model to extract and model the emotional state of an agent and use such information in the RSA implementation of irony modeling to improve the communicative model's performance.
- **Chapter Five** contains the conclusions, a description of what can be improved or be done differently, the weak spots of our approach and how to fix them, and possible future developments.

- **Appendices** Appendix A contains a definition of probabilistic programming. This methodology is the basis of the Rational Speech Act framework and the Variational Autoencoder, the cornerstone models of this thesis. Appendix B describes the main technologies and external tools used in implementing our project. Appendix C contains a detailed exposition of the MVAE models' implementation we utilized in our project.

Chapter 2

Concerning emotions

Through the years, philosophers, psychologists, and neuroscientists have described emotions through numerous definitions, and many theories about their function, evolution, and nature, in general, have been proposed.

In this chapter, we will explore the main theories of emotions and how they can be adopted in computer science, declining each definition of emotion into a computable form to be utilized in affective computing programs.

It may sound counterintuitive at first, but emotions have been proven fundamental for rational thinking. According to neuroscientist Antonio Damasio, emotions play an essential role in social functioning decision-making, and learning processes [5]. Citing one of Damasio's studies,

> "... decision making is a process that is influenced by marker signals that arise in bioregulatory processes, including those that express themselves in emotions and feelings." [6]

Furthermore, in the book Descarte's Error [7], he states:

> "The action of biological drives, body states and emotions may be an indispensable foundation for rationality. The lower levels in the neural edifice of reason are the same that regulate the processing of emotions and feelings, along with global functions of the body proper such that the organism can survive."

According to this vision, emotions arise due to our bodies' regulatory processes to grant survival. Damasio theorizes that emotions operate as a guide in the decision-making process. He defines the "somatic marker" as internal feedback that marks experiences and associates them with body sensations such as increased heart rate or visceral responses. Those interoceptive feedbacks would help the reasoning and decision-making when subjected to similar circumstances. Emotions are anchors that

define and guide our reasoning process, binding our "rational mind" to our bodies' internal states and feelings. Emotions are a cornerstone in how we process and interact with the world and are fundamental to pragmatic reasoning.

By being studied in different fields of knowledge, theories on emotions vary significantly from each other. In order to make preliminary coarse-grained discrimination between them, we borrow the wave-particle duality concept from quantum mechanics, where every particle may also be described as a wave; we can find a similitude with the case of emotions.

According to some scholars, emotions are better described as distinct and discrete phenomena, both regarding other mental capabilities and emotions themselves; an independent faculty characterized by its process separated from the others through ad-hoc mechanisms and neural "circuits." Some examples of this school of thought are the Basic Emotion Theory (or BET) and the casual appraisal theories. Others, in turn, prefer to describe emotions as a complex set of phenomena better described as a multi-dimensional space of different continuous variables. Individual "discrete" emotions appear as the categorization of a series of biological, physiological, and external circumstances, a population of highly variable instances of events with a certain degree of similarity between them. The theories supporting this view are the constructionist theories of emotion.

If we observe the matter from this level of abstraction, the following quote from Albert Einstein appears to fit our case: "It seems as though we must sometimes use the one theory and sometimes the other, while at times we may use either. We are faced with a new kind of difficulty. We have two contradictory pictures of reality; separately neither of them fully explains the phenomena of light, but together they do." Of course, Einstein was referring to a different matter, but if we substitute the "light" with "emotions," the similitude can hold.

Luckily, it is not our aim to inquire which vision of the subject is the one that depicts reality more precisely and, in some respects, a more intuitive high-level interpretation of the phenomena could better suit our issues. In the following chapters of this thesis, we will see how the communicative act between people can successfully be described and emulated through a recursive model where each actor can understand the other thanks to its ability to emulate its affective state. In simpler terms, by putting themselves in the shoes of their listener, a speaker can decide the best sentence to utter in order to achieve their goal by reasoning as if they were the one to whom the sentence referred.

This process requires an intuitive knowledge of emotions, their cause, and their consequences. As one can predict the trajectory of a ball thrown in the air by only having an intuitive knowledge of physics behind it, one person can predict the consequence of insulting another person, and knowing that, they may choose not to. We define lay theories of emotions [8] those theories that rely on people's intuitive

knowledge of emotions and the ability to reason about them (affective cognition) to model social behaviors and communicative acts.

In the last two decades, theories of emotions found fertile soil in computer science giving birth to the interdisciplinary field of affective computing. Affective computing involves computer science, psychology, neuroscience, philosophy, art, and industry in the effort to build techniques and technologies that convey, recognize or influence emotions. Let us now delve into the main theories we broadly describe early in this section in order to highlight how some of those might find practical use in the field of affective computing.

2.1 Basic Emotions Theory

Basic Emotion Theory (BET) represents the classical view of emotions as discrete entities; its most notorious exponent, Paul Ekman, theorizes that each emotion manifestation occurs through specific physiological and behavioral patterns, such as facial expressions.

Ekman leads the origins of his theory back to Darwin's studies of evolution. Facial movements are considered adaptations used to communicate internal states such as distress or affection or similar conditions, an essential feature among the group-living species. Darwin listed a first collection of discrete emotional states communicated through facial expressions composed of pleasure, joy, affection, pain, anger, astonishment, and terror.

According to Ekman, emotions are triggered and executed through a "facial affect program" (FAP), a set of specialized central neural structures that directly map each discrete emotion to its corresponding facial pattern.

To summarize, some key points behind the basic emotion theory that distinguishes it from other theories are that emotions are unique mental states caused by ad-hoc mechanisms and "cabled" in specific brain circuits. Each emotion has a unique and distinguishable external manifestation (such as facial expression) and specific internal body states. Emotions are universal and pan-cultural. Each emotion evolved individually.

Being the activation of facial muscles unintentional and cabled through special brain circuits, facial expressions are considered a reliable source of information about the internal emotional state of the individual. This aspect led to the development of a method for measuring and evaluating facial movements. The Facial Action Coding System (or FACS)[9] is a methodology that encodes facial expressions through the mapping of the state of specific muscles or groups of muscles, called action units (AUs). By mapping every single component that forms a facial expression through the AUs and under the theorized assumption that each facial expression is uniquely

associated with emotion, the FACS approach promises to identify and code each discrete emotion.

Although it is generally accepted that facial expressions are a good source of information for the affective state of the individual and hence a valuable tool in communication, some critics of the Facial Action Coding System have risen since the proposal of the method. Barret et al. [10] criticized some of the strong assumptions of the system, such as the reliability of the emotions categories and the lack of specificity, meaning that the unique mapping between specific facial configurations and instances of the same emotion category is not consistent, and the same emotion category might be expressed through different facial configurations and vice versa. Other critics are directed toward the assumption of the basic emotions' universality since the impact of context and culture have not been sufficiently documented. Although the theorists of the basic emotion theory have accepted part of the criticism by weakening some of the strongest assumptions, the controversy toward Ekman's work has not ceased, especially in light of advances in neuroscience.

2.2 Constructionist theories

On the other side of the spectrum, with respect to Basic Emotion Theory, we have the constructionist view of emotion. It includes a set of heterogeneous theories with, at the core, the idea that emotions are not special mental states with ad-hoc neural circuits, each with a unique purpose or manifestation. Constructionist theories might not be aligned in the designation of the phenomenon's origin; according to psychological models, emotions are roughly defined and highly variable categorization of a continuous constructive mental process composed of a multitude of mental elements. On the other hand, social construction models see the source and characterization of emotions in the social context. According to this view, emotions are not a mental construct as much as the fruit of culture. The mental components of emotions are seen as a function of social meanings and are considered mainly from the perspective of their social function.

Our attention will mainly focus on the first type of constructionist theories, hence the psychological construction models. In particular, the work of James Russel [11] provides us with the theoretical background to build a computational model to represent emotions. Russel's studies depict a continuous emotional mental state which emerges from the activity of multiple stimuli and perceived feelings. This space, called Core Affect, is the integration of two components, valence and arousal. Arousal is the level of activation, the intensity of the feeling in a given moment Valence represents a pleasure-displeasure range of possible values that describe whether a sensation has a positive or a negative connotation for the person that feels it. The concept of categorical/basic emotions we have encountered in the BET model is now substituted

with the idea of prototypes. Unlike what has been said for basic emotion, prototypes are secondary concepts that emerge from the core affect and are also defined by the external context. [12]

2.3 Emotion representation in computer science

In this chapter, we introduced some of the more representative theories of emotion and identified some key concepts that distinguish and characterize them. As we anticipated, our intent was not to provide an exhaustive and complete list of the history and development of the studies on emotions but to identify those theories that are both well accepted by the scientific community (even if not necessarily universally accepted) and, at the same time, are well suited to be used as groundings stones for computer science applications. Until now, we proposed two ways of interpreting the concept of emotions and the phenomenon around them: we introduced the BET view on the topic that considers each emotion a distinct and independent process, both from the perspective of its manifestation and biological/neurological source. Next, we described the constructivist proposal of James Russell, where the emotions category concept appears as a secondary element emerging from what is at the center of the affective phenomenon, the Core Affect, a mental space formed by the merging of several interoceptive biological feedbacks that adds up in the dimensions of valence and arousal.

Through the years, those two theories have gained the attention of researchers in the AC field, where most of the effort is spent on the task of emotion recognition through the classic pattern recognition approach. This approach uses machine learning techniques to extract features from collected data to identify and classify patterns. The data sources may consist of sensor recordings of biological signals depending on the background theory of the study. The Basic Emotion Theory fits this kind of approach since, according to this view, emotions are of finite number, clearly distinguishable, and univocally identifiable from external features such as facial expressions. Given those premises, it is easy to build machine learning models under the supervised learning paradigm; labeled data points (usually through human annotation) are submitted to the model under training, which tries to classify them and learn by comparing its prediction to the ground truth provided by the label. Reasonable data sources for this method may be represented by facial expressions, voice recordings, or other physiological signals to whom it is possible to associate a categorical emotion.

Similarly, applying such an approach to the constructivist theory of the Core Affect is also possible. The main differences from the previous case consist of the different labeling; in this instance, we associate each data point with two continuous values (variance and arousal) instead of a categorical one.

By following this approach, it is also possible to train models to generate a mapping between categorical emotions and the V/A space. Such translation between different theories of emotions is interesting from an operative perspective. On the plus side, the mapping between categories and affective space allows a straightforward interpretation of the core affect, even if it tends to simplify its interpretation. At the same time, it allows for a representation on a continuous scale of discrete values, which is helpful for specific applications. It is worth noting that the theories do not support this comparison since emotions and affective states are not synonyms and represent different concepts according to the appraisal theories.

The issue of representing the ensemble of interoceptive stimuli perceived as the affective state with a categorical instance, even if there are reasons to criticize Ekman's approach, is not negligible. In order to communicate, people need to create categories representing fuzzy concepts. Generally speaking, categorization is the act of grouping similar objects together according to some features they share. We can define perceptual categories when we use perceivable physical features of the objects in order to perform the grouping.

We define abstract categories when we group objects or instances by their functional similarity in different situations. For example, even if an airplane and a motorbike are not comparable in shape, both may fall under the same category of means of transport. In the latter case, they are grouped not by their shape but by their shared purpose.

It is also possible to distinguish multiple abstract categories in the same perceptual category. Even though both share four wheels, a motor, a steering wheel, a Formula 1 car, an off-road car, and a city car are grouped in different abstract categories since they serve different purposes.

Dr. Feldman Barrett [13] tackles the problem of defining the type of category the emotion fell into and defines them as abstract ad-hoc categories. Ad-hoc categories are context-dependent and are subjected to high variability in their manifestation patterns while maintaining a similar (but not identical) functionality. Each emotion category's purpose depends on the context and situation and manifests in the "most appropriate" way given the situation and its purpose. This association between emotion manifestation, purpose, and context is learned during the growth of the individual and depends on the cultural environment.

In light of what is said above, it is clear that in order to interpret emotional concepts correctly, it is necessary to take into account several ingredients: the association between the affective state; the current context; and the prior knowledge that associates each context to the appropriate response and purpose. We could add a word (such as anger, fear, happiness) to represent the concept that results from it in natural language.

It appears clear how we need a more structured model of emotion to correctly

understand its phenomenon and better employ it in our applications. The Conceptual Act Theory seems to be able to furnish a theoretical basis that would guide us to a better computational representation and use of emotions.

2.4 Conceptual Act Theory

The Conceptual Act Theory [14] sees emotions as mental constructs generated by the brain in the moment of need. Emotions are constructed concepts associated with labels such as "happiness" or "fear" that result from the interpretation of an occurring event under the light of previous experiences. [1]

Nevertheless, emotions are not the only concepts our brain generates. Emotions arise from the action of labeling the interoceptive experiences of the body as the set of sensory perceptions originating from the constant change of physiological states of regulatory processes of the organism. At the same time, the human brain builds concepts of exteroceptive perception, such as visual, auditory, or tactile inputs. All those concepts are elaborated together to generate the whole perceived world the individual is absorbed.

All those concepts are not "built in" at birth but are learned with the individual's day-to-day experience. Links between emotion abstract categories and labels are forged according to a mechanism of learning by example, in which the individual associates words with concepts according to what he observes from others.

Language plays a fundamental role in forming concepts and, consequently, in forming, elaborating, and communicating emotions. If the prior knowledge and the concepts are similar enough, and the used labels are the same, it is possible to transfer desired meaning through the communicative act between persons. In other words, communication is possible (or at least is eased) if the individuals speak the same language, associating the same concepts/experiences with the same labels/words.

In turn, emotions provide an essential tool during the communication act; conveying affect-based fragments of information through language or non-verbal behaviors enriches and simplifies the expression of intentions, needs, and desires.

According to this vision, the brain forms, exchanges, and elaborates concepts to maintain and preserve the body's physiological balance, the allostasis, by performing simulations and predictions of how the external and internal conditions might alter and adapt to them. This behavior defines the Bayesian brain as a probability machine: it forms predictions about the world using the knowledge formed through

[1] Note: category and concepts should not be confused: a category represents a set of events of objects with similar features or that serve a similar goal according to the context; a concept is the set of representations that corresponds to those events of objects. In the case of emotions, the concepts are the mental representation of the emotion category

past experiences; performs actions that would change the condition in favor of the organism; and then updates its knowledge and prediction based on what it receives from the sense as feedback from the mutated environment.

In the described process of evaluating the world's state through the information that comes from prior knowledge, emotions serve the role of the compass that indicates the aspects of the current situation which are more relevant and need attention filtering out all the rest.

The process of evaluating the external situation through sensory inputs and integrating the physiological and bioregulatory systems of the body is enabled by the psychological primitive we formerly introduced under the name of core affect. Prior knowledge enables the interpretation of the feelings that arise from the synthesis operated by the core affect, and concepts such as emotions are the result of this mapping.

To summarize, Categorical Act Theory considers different psychological moments, each of which addresses different mental capabilities:

- We define perception as the ability of the mind to focus on those sensations that derive from inputs of the external world.
- We define cognition as the psychological moment when the mind focuses on elaborating, through the use of prior experience, mental concepts which might refer to previous moments (memory), present moments (thinking), or future events (imagining).
- We define emotion as the mind's ability to focus, understand and elaborate on the sensations that arise from the body.

According to those psychological moments, the brain defines strategies and actions to alter the state of the world. The consequences of those actions are then perceived by the individual, which changes his internal state and prior knowledge accordingly.

In this chapter, we posed the theoretical basis of using emotions as a tool to elaborate and understand the concepts our mind builds. We proposed a series of theories (Ekman's BET, Russell's Core Affect) that could, and have been, used as the basis for affective computing applications of emotions. Ultimately, we described a more complex view on emotions, the Categorical Act Theory, that inserts emotions inside an elaborate framework that aims to describe how our minds build and elaborate concepts to interact with the surrounding environment. In the next chapter, we describe a framework, the Rational Speech Act, that tries to formalize pragmatic reasoning and inherit some aspects of the Categorical Act Theory.

Chapter 3

Rational Speech Act

This chapter explores the Rational Speech Act framework from a theoretical point of view. The framework inherits from Grice's Cooperative principle to shape the recursive behavior of the agents involved, go beyond the literal meaning of the words, and interpret the language in context. The Rational Speech Act (by Goodman and Frank, [1]) is an agent-based framework that formalizes pragmatic reasoning by drawing on theories derived from neurobiology, psychology, and the philosophy of language.

RSA borrows some intuitions from Grice's Cooperative Principles: the assumption that two agents involved in the communicative act are cooperative and that the speaker respects the principles of Quantity, Quality, Relation, and Manner:

- Quantity - Avoid obscurity: (i) Make your contribution as informative as is required (for the current purpose of this exchange); (ii) Do not make your contribution more informative than is required.
- Quality - Try to make your contribution one that is true: (i) Do not say what you believe is false; (ii) Do not say that for which you lack adequate evidence.
- Be relevant.
- Be perspicuous. (i) Avoid obscurity of expression; (ii) Avoid ambiguity; (iii) Be brief; (iv) Be orderly. Not all maxims, however, can be respected in all circumstances, such as the cases of ironic statements, metaphors, or understatements. In those cases, the Quality principle is obviously not fulfilled.

The RSA framework tries nonetheless to furnish a method to overcome the issues brought by the need for implicature, allowing to "solve" non-fully-determinable cases. According to Grice's theory, the Speaker chooses his utterance and actions to be intelligible, and the Listener operates according to this assumption. Regardless of the Speaker's particular goal, the primary objective is always to be understood by

the other actor. The recursive reasoning between the actors involved requires a certain similarity in the forms of reasoning of the participants of the communicative act. The "similarity" principle means the agents can simulate how the other would reason and react to their actions and the shared environment, allowing both to establish the best strategy to reach their goal. This principle assumes they share a similar internal neurobiological and psychological structure, the same external context, and similar prior knowledge. The ability to "think what the other may think" enables the recursive reasoning approach, which is the mechanism at the core of the framework.

Regarding Grice's theory, RSA adds the element of emotions and feelings. It adopts the perspective of the Conceptual Act Theory to model emotions as concepts generated from integrating physical sensations (core affect) with previous knowledge, given the current circumstances.

The elements/concepts modeled by the framework may vary in number, complexity, and relations according to the type of interaction being simulated. In our case, we want to model an interaction that involves irony, and some of the elements that might occur are the following:

- Goals: represents the aspects the Speaker wants to inform the Listener about. It relates to its emotional state concerning some aspects of the current shared external context.

- Utterance: The sequence of words the Speaker emits during the speech act, and the Listener hears. The words used are linked to the concepts and beliefs shared between the Listener and the Speaker. They belong to a shared lexicon over which they are chosen, and their expression as a speech act represents the action the Speaker chooses to reach their goal.

- Language Model: The language model describes the rules that form a language according to a probabilistic view. In its most general sense, it uses statistics to determine the probability of a sentence being formed. The actors can use it to generate correctly formed sentences.

- Meaning: The meaning is what the speaker desires to communicate through the uttered words to the Listener. The meaning may be linked directly to the spoken sentence and adhere to the exact significance of the words (base case, the literal meaning). It may refer to something not explicitly expressed but left to the Listener to imply (implicature). The literal meaning of the expressed sentence could imply something absurd or out of scale for the current context. In contrast, the desired meaning the Speaker wants to share may only accentuate or describe an aspect of the selected objective (hyperbole). The literal meaning could imply the exact opposite of the concept the speaker desires to communicate; this last case is the case of irony.

- Beliefs: The beliefs represent the prior knowledge (of both the Speaker and Listener) about words, feelings, desires, and goals and link and connect all those concepts. They might regard elements from the outside world (through exteroceptive sensory systems) and feelings and sensations from inside the body (through interoceptive perception).

- Actions: are what the agent does to engage in the communicative act. For example, the actions of uttering a sequence of words, activating the facial muscles to express emotion, or taking a specific physical pose by moving the whole body around the space. In all those actions, one agent might decide to convey a specific meaning coherently with the current context and its prior beliefs.

- Observation: Represents the union of all forms of information at a given time. It comprehends the observable results of previous actions, the current state of the context/world, and the perceivable information from the speaker/listener. Affective State: Represents the interpretation of the body's internal status as provided by its interoceptive feelings. It is expressed through the variable of Valence and Arousal and is bound to the conceptualization of categorical emotions.

- Concepts: Concepts are mental instances of external categories. They enable the relations between language, beliefs, core affect, and external inputs. For example, we can build the concepts of discrete emotions that link together language elements, such as the words "happiness", "sadness", and "fear", with the corresponding affective state (V/A) according to the current scenario in which the subjects are experiencing.

Those elements' interaction rules are derived from psychological and neurobiological theories. However, we model them through probabilistic dependencies that enable us to infer the internal state of one agent (their internal unobserved random variables), given the observable variables and the knowledge of the rules and structures that form the model.

We will now describe some of those possible relations and interactions, focussing our attention on the case of the representation of irony in the communicative act. It is worth noting that each structure and element we describe could be easily modified or integrated with additional modules or links whenever supported by new theoretical work.

3.1 General model

In its general formulation, the Rational Speech Act involves the interaction between two agents: the Speaker and the Listener. Both agents reason recursively about how the other would reason and use this knowledge accordingly: Either to perform the specific action that would lead to the desired outcome. In the case of the Speaker, this determines the choice of the utterance u. Alternatively, in the case of the Listener, evaluate the performed action and come to the correct conclusions about the counterpart's intentions and goals.

The recursive structure of the framework requires a base case; otherwise, it would never reach an end. The base case of the recursion is interpreting the Speaker's statement according to a Literal Listener. The Literal Listener will only assess whether or not the Listener's utterance u corresponds to a possible given meaning m according to its literal semantic:

$$P_{Lit}(m \mid u) \propto \delta_{[[u]](m)} P(m) \tag{1}$$

Where [[u]] is a denotes for each sentence whether or not the utterance u is true of a given meaning m, and $P(m)$ is the prior probability associated with the given meaning.

The Speaker who knows how the Literal Listener reasons select the utterance u to convey the desired message:

$$P_S(u \mid m) \propto \exp \alpha \log P_{Lit}(m \mid u) \tag{2}$$

where $\alpha \in [0, \infty)$ parameter controls how rational the speaker is. In turn, the Pragmatic Listener reason recursively about the Speaker's utterance, knowing the model that led to the choice of the spoken sentence.

$$P_L(m \mid u) \propto P_S(u \mid m) P(m) \tag{3}$$

3.2 Shaping Irony

Irony is a complex form of speech where the literal meaning and the real meaning of the spoken utterance are opposite. To discern a sentence's real significance is insufficient to rely only on the spoken words. It is essential to consider the context, the non-linguistic sources of information, and the prior knowledge shared by the agents involved to go behind what is said and understand what is meant. It is required to make use of pragmatics.

In their paper [15], Kao and Goodman propose an application of the Rational Speech Act framework to model Irony. We will use this application as a starting

point to build a more complex model, which also considers using facial expressions as a cue for the Listener.

Let us describe the case of irony from a high level of abstraction: We consider the situation where the two agents, the Speaker, and the Listener, are together on a tropical beach on a sunny and cloudless day. This scenario represents the context that both the agents share.

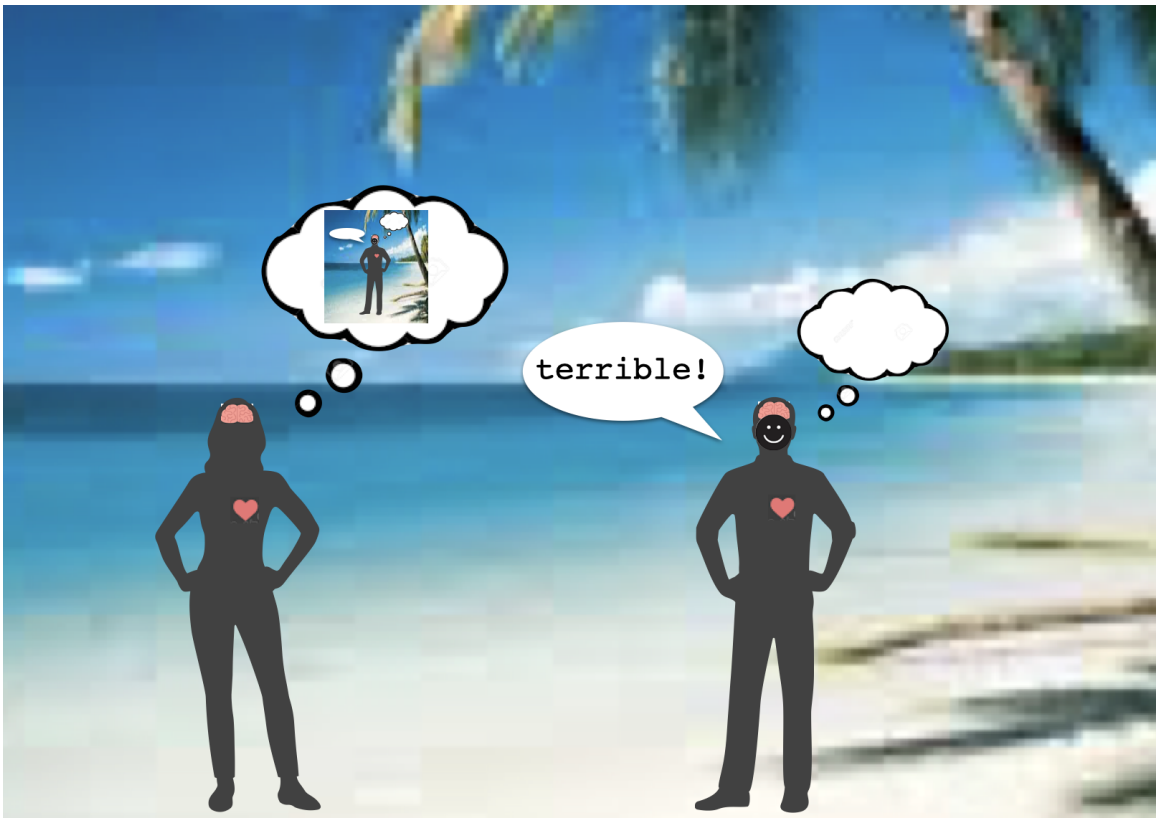

Figure 1: Image representing the scenario of our analysis. The context is a beautiful tropical beach on a sunny day. The Speaker utters the word "Terrible" and smiles as a result of a non-observable internal state; the Listener sees the smile and hears the utterance and recursively reasons about the Speaker's actions in order to assess his internal state and understand the true meaning of the word given the context.

The context is paired in the mind of both Speaker and Listener with pre-existing beliefs: For example, if the sun is in the sky with no clouds, the weather is considered "good". Good weather is, in turn, associated with positive feelings (emotional state), and words such as "Amazing", which in turn, as associated with positive emotional concepts.

We see how the state of the world is associated with the emotional state (core affect) through a shared set of beliefs and concepts, which are associated with a set of words.

Both Speaker and Listener have a shared notion of what the common beliefs about the current state of the world might be. Consequently, they share prior knowledge of the most probable internal state of one another at the emotional level (core affect) and the concept and words associated with such an emotional state.

The Speaker can therefore choose his utterance by considering the state of the world and the prior beliefs shared with the Listener about it. In the case of irony, the Speaker will probably choose a statement that subverts prior knowledge about one aspect of the context (or the various concepts, beliefs, or emotions associated with it).

For example, let us assume that the Speaker states the word "Terrible": The Speaker selects an aspect of the state of the world or his emotional feeling about it. He then selects the utterance by reasoning recursively about what sequence of words would be received as unexpected by the Listener. That sentence, whether considered true by the Listener, would define a context, concept, or emotional state that contradicts the prior knowledge of the Listener about the current shared state. The Listener, in turn, will reason recursively about the Speaker's word choice, considering what could be the element they are referring to and what is the actual emotional state of the Speaker towards those aspects.

Let us now formalize those elements and their interaction: First, it is necessary to define the question under discussion (QUD). Complex figures of speech, such as irony, tend to be expressed through short utterances, and most of the context remains inferred through other non-literal means. To infer the subject of interest, we define the set of possibilities the Listener has at her disposal inferred by the current state of the world. We describe the state of the world as s, as the entirety of the possible subjects of interest the Speaker may refer to, given the current context shared by both agents. We also define A as the Speaker's emotional state towards one aspect of said s.

The inference of the correct meaning requires that the Speaker is cooperative and informative about the QUD. Furthermore, the RSA framework assumes that speaker and listener reason recursively about each other. It follows that the Speaker chooses the utterance that best conveys information about the QUD by maximizing a utility function that a literal listener (a hypothetical listener that assumes the Speaker's utterance is always true) would also know and utilize.

According to the utility function U, we assume that the speaker S chooses the utterance through a softmax decision rule that rewards the most informative sentences considering the literal Listener's L_{lit} background knowledge.

$$S(u|s, A, q) \propto e^{\lambda U(u|s,A,q)} \tag{4}$$

$$U(u|s, A, q) = log \sum_{s',A'} \delta_{q(s,A)=q(s',A')} L_{literal}(s', A'|u) \tag{5}$$

Where u represents the spoken utterance, s the state of the world, A the affective state of the listener about s, q the QUD, λ is an optimality parameter, and $L_{literal}(s', A'|u)$ represents the literal listener updating her prior beliefs, given u be considered true.

As we previously stated, the QUD itself needs to be inferred: as a consequence, the Pragmatic Listener ($L_{pragmatic}$), in turn, will reason recursively about the Speaker's utterance to marginalize the possible QUD:

$$Lpragmatic(s, A|u) \propto P(s)P(A|s) \sum_{q} P(q)S(u|s, A, q) \tag{6}$$

In this base case, the new information introduced by the Speaker and over which the Listener performs inference is only in the form of spoken words.

We now want to provide the Listener with an additional source of information by extracting the values of A from the Speaker's facial expression.

In our scenario, the utterance represents an action performed by the Speaker. We introduce a second action: the Speaker's facial expression in the instant of the utterance.

This second action allows the Listener to integrate her prior knowledge about the Speaker's emotional state with new information. This addition is supposed to affect the recursive inference process performed by the Listener, allowing a more accurate depiction of the internal emotional state of the Speaker and the meaning behind his words.

This chapter introduced the Rational Speech Act, a framework to model and implement pragmatic reasoning. We listed and described some elements that concur and enable the communicative act for the non-trivial cases of ironic statements.

In the next chapter, we will delve into implementing the RSA model for irony to demonstrate the influence that the introduction of emotions operates in the communicative act as a positive contributor to implicature. We will also propose a method to generate and interpret emotion through facial expressions using a Multimodal Variational Autoencoder model. Such a model would allow us to represent the additional action performed by the Speaker in the version of our RSA model that includes the interpretation of the affective state through the facial expression.

Chapter 4

Implementation models

In this chapter, we describe the choices, the issues, and the results of our project's implementation. We refer to the appendices for further details on the low-level implementations or technology.

To briefly recall the objectives of this project: we intend to study how pragmatic interpretation of communicative acts is facilitated by using emotions as a vehicle of information, allowing the discerning of complex figures of speech that conceals a true meaning behind the spoken words.

In the last chapter, we presented the Rational Speech Act framework as a tool to implement pragmatic reasoning, go "beyond what is said," and comprehend the true meaning of an utterance considering the context. RSA models the individual's affective state as a tool to indicate where the attention of the agents involved in the speech is focused on and which aspects of the current state of the world are the subject of the communicative act.

To approach our goal, we divide the project into two parts:

- One part will focus on implementing a system that would allow us to handle the generation and evaluation of the affective information derived from facial expressions. The developed tool should be able to generate and recognize facial expressions and associate them with correct affective state values and categorical emotions.

- The second half of our project concerns the implementation of an instance of the RSA framework that focuses on discerning ironic statements.

Finally, we apply the tool we developed in the first part of the project to the RSA model in order to generate the facial expression of the Speaker based on his affective state and provide the Listener with an additional source of information to perform inference of the meaning behind the Speaker's actions. We then evaluate whether introducing emotions as an additional source of information improves the model's

performance concerning the simpler instance of the RSA, which does not consider facial expressions.

4.1 Emotions Representation Through MVAE

In this section, we describe how we handled the affective components of our project through the generation and evaluation of facial expressions. Our goal for this part of the project is the realization of a model that would allow generating facial expressions associated with categorical emotions and, vice versa, recognize the affective state behind a facial expression.

We approach our problem by using Multimodal Variational Autoencoders. Autoencoders allow us to encode its input into a latent space and retrieve it via a decoder. They are often used for dimensionality reduction; hence it maps higher-dimensional data points into a lower-dimensional space.

Variational autoencoders (VAE) model the latent space as a probability distribution, which allows the sampling of new instances of latent vectors to generate new representations of data points through the decoders. This introduction of the sampling step makes the whole process non-deterministic; hence it allows for a probabilistic interpretation of the problem.

Using variational autoencoders, we aim to build a model able to produce latent representations of emotional concepts and use those to generate new instances of categorical emotions and visual representations of emotions through facial expressions. To mix those two forms of representation of the same concept, we need methods to encode two different kinds of input into the same latent space and two decoders that generate two different representations of the same emotional latent concept.

However, let us take a step back to formalize the Variational Autoencoder that models through a Bayesian framework. We define a latent variable model in the form of:

$$p(x, z) = p(x|z)p_(z) \tag{7}$$

Where x is our observed variable belonging to our data set, and z is the latent representation of the observed variable. We aim to implement a function by learning a set of parameters θ in order to approximate the distribution p.

We want to obtain the posterior distribution $p(z|x)$ where z is a random variable in the latent space, and x is input data. The posterior distribution would allow us to map each input data point x into its latent representation z. We also want to be able to perform marginal inference over x, given a latent variable z, hence model $p(x|z)$.

According to the Bayes theorem:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{8}$$

Our main problem is that $p(z|x)$ is hard to compute due to the unknown evidence distribution $p(x)$; for this reason, instead of the true posterior $p_\theta(z|x)$, we are going to use a parametrized distribution $q_\phi(z|x)$, where ϕ represents the sets of learnable parameters that would allow us to approximate the objective distribution. In our instance, q represents the encoder(s) in our VAE.

We train the variational autoencoder by maximizing the Evidence Lower Bound (ELBO).

$$logp(x) \geq \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - KL(q_\phi(z|x)||p(z)) \tag{9}$$

Where $q_\phi(z|x)$ is the "proxy" distribution we are trying to learn in order to approximate the true posterior $p_\theta(z|x)$. It is our encoder used to generate new samples z of the latent space. In Equation 9, we compute the Kullback–Leibler divergence between $q_\phi(z|x)$ and $p(z)$;

$p(z)$ is the prior of the model we consider to be a fixed unit Normal distribution; It serves the purpose of enforcing a gaussian shape to the distribution $q_\phi(z|x)$, allowing it to learn more meaningful representations. Finally, we have the $logp_\phi(x|z)$term; it is the log-likelihood of the observed x given the sampled z; the decoder implements it in the VAE model, also parametrized in θ. High log-likelihood values are associated with the real data being correctly reconstructed starting from latent space samples.

The encoder's output is represented by two parameters that define the normal distribution, hence the mean (μ) and the variance (σ^2). We use those values when generating new samples of the latent space from our encoder. The decoder is the neural network that allows the mapping from the latent space to the input data and implements the distribution $p_\theta(x|z)$.

Our goal during the training is to tweak the network parameters to approximate the true posterior distribution. We use the Kullback–Leibler divergence to measure the distance between the parametric posterior distribution (defined by the parameters of the neural networks) and the objective posterior. This process allows the autoencoder to organize the latent space in such a way as to generate meaningful samples and reduce the possible overfitting of the model. The multimodal version of the Variational Autoencoder adds an ulterior step to the process; it uses different encoders and decoders to map different inputs into the same latent space by combining the probability distributions generated by each encoder through the product of experts [16] technique. See Appendix B for more details about product of experts and other modalities merging techniques.

In their work [17] Zhao, Song, and Ermon address the problem of the training objectives for the variational autoencoders; they affirm that in case of limited resources, the ELBO objective favors fitting (and eventually overfitting) the data distribution over performing correct inference.

We encountered an issue during the first attempts at the model's network training that could be traced back to the issue discussed in the cited paper. In our case, the model training reached a state of convergence of the loss function. However, all the image samples represented the same blurred depiction of a human face, where few to none of the somatic traits of the individual were recognizable, regardless of which categorical value the sample was conditioned over. We theorize that the high complexity of the decoders caused the model to ignore the sampled vectors. Instead, it learned a set of parameters that would allow the network to generate images good enough to produce a low enough reconstruction error, but that did not correctly represent the target distribution. As addressed by the paper's authors, the problem was partially solved by reducing the size of the neural networks (in particular, the decoders) and by introducing modifications to the loss function as proposed in the cited paper.

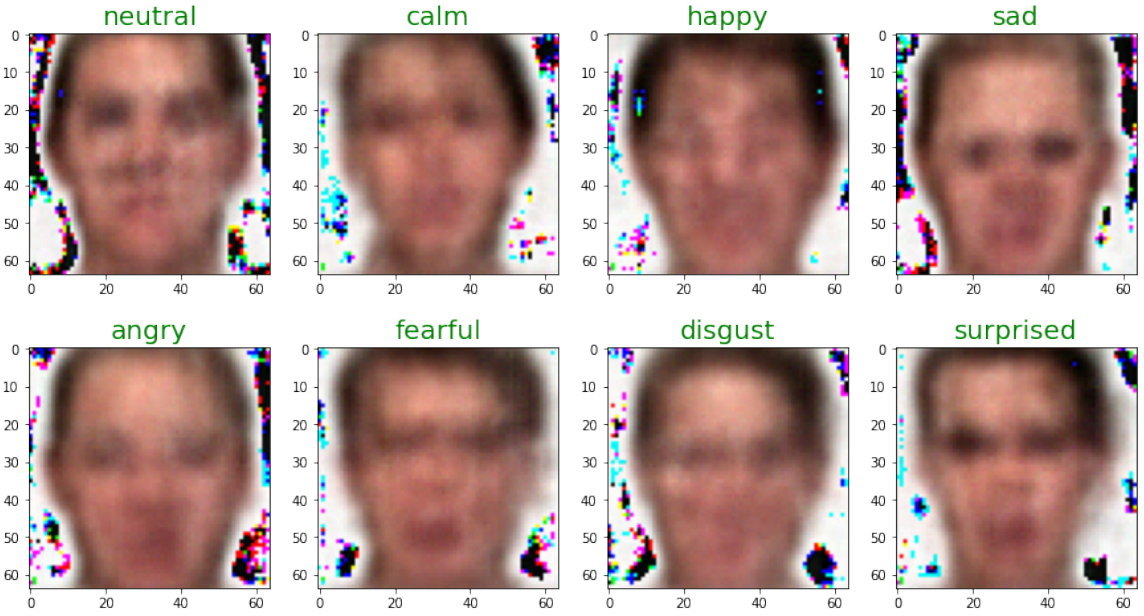

Figure 2: Examples of images generated by decoding the latent space samples when directly using raw images as a modality. Each image is sampled from a latent vector after encoding a different categorical emotion.

Zhao, Song, and Ermon propose to modify the ELBO training function as follows:

$$\begin{aligned} LInfoVAE = \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - \\ (1-\alpha)\mathbb{E}_{pD(x)}D_{KL}(q_\phi(z|x)||p(z)) - \\ (\alpha+\lambda-1)D_{KL}(q_\phi(z)||p(z)) \end{aligned} \tag{10}$$

The Info-VAE loss function is the general case that comprehends the previous models: the first and second elements of the formula are the same that appear on the standard ELBO loss function, namely the reconstruction loss and the KL divergence loss. The Info-VAE is equivalent to the ELBO loss function for the values of $\alpha = 0$ and $\lambda = 1$. The β-VAE we previously we mentioned corresponds to the Info-VAE when $\lambda > 0$ and $\alpha + \lambda - 1 = 0$. We notice then, that the main change is represented

by the introduction of the new term $(\alpha + \lambda - 1)D_{KL}(q_\phi(z)||p(z))$.

The ratio behind the introduction of this additional constraint is to lead the distribution of the possible vectors generated by the encoder $q_\phi(z)$ to assume a similar shape to the prior distribution of possible well-formed vectors. This prior distribution $p(z)$ is usually chosen as a simple normal distribution. The Kullback–Leibler divergence term present in the "standard" ELBO formulation, $KL(q_\phi(z|x)||p(z))$, tends to guide the model into generating for each input point x samples of the latent vector z matching the prior distribution p(z), "averaging" the output vectors regardless the input. The new divergence term is, however, hard to compute but can be replaced with another strict divergence, defined as $D(q\phi(z)||p(z)) = 0$ if $q\phi(z) = p(z)$.

We choose to implement one of the divergences proposed by the paper mentioned above, the Maximum-Mean Discrepancy (MMD):

$$D_{MMD}(q||p) = \mathbb{E}_{p(z),p(z')}[k(z,z')] - 2\mathbb{E}_{q(z),p(z')}[k(z,z')] + \mathbb{E}_{q(z),q(z')}[k(z,z')] \tag{11}$$

MMD is a divergence that measures the differences between two distributions through their moments. It can be implemented through the "kernel trick," a method that compares the differences between two samples of the distributions of interest.

The described model allows us to adapt the input-output data domain according to our necessities. The modularity of the approach would also permit the use of additional modalities other than the image-category couple we adopted.

In the context of the communicative act, we could have adopted other sources of affective information, such as the prosody of the speech or any other cue to the individual's affective state. For each additional modality, the model would have required an additional encoder that would map the input to the latent space and a decoder to sample new instances of the selected modality from the latent space.

For what concerns our work, we implemented two models to map emotion categories and facial expressions. In our first implementation, we used raw images depicting men and women involved in a speech act simulating facial expressions labeled with the relative categorical emotion. For our second implementation, we first extracted the "affective features" expressed by the actors through facial cues in the form of Action Units by using the external tool OpenFACE (see Appendix B for details).

Action units are a way to describe the configuration of facial muscles by mapping their level of activation through numerical values. We used the extracted features as an input modality for our model instead of the raw images. In order to revert from AUs to the facial expression, we adopted a second external tool, OpenFACS, that, given the values of the Action Units, it would generate a 3D model representing a human face in the described facial configuration.

We refer the reader's attention to the Appendix C for a more in-depth description of the models' implementation, training, and performance evaluation.

(a) angry (b) calm (c) disgust (d) fearful

(e) happy (f) neutral (g) sad (h) surprise

Figure 4: Examples of images generated by decoding the latent space samples when directly using Action Units as a modality. Each image is generated using the OpenFACS software by submitting the decoded AUs values after passing categorical emotions as input values to the model's encoder.

4.2 Rational Speech Act Implementation: Irony

In this section, we describe the implementation details of the rational speech act model applied to irony. Irony is a complex form of speech and is described by the Cambridge Dictionary as: "a form of deliberate mockery in which one says the opposite of what is obviously true."

We consider a first implementation of the RSA framework in which we want to test the ability of the model to correctly understand the depicted situation, in which two human actors are involved in a speech interaction, and the Speaker utters an ironic statement about their shared context.

Once we have defined the base model, we introduce an additional source of information for the Listener in the form of emotions expressed by the Speaker through facial expressions. Emotions are possibly generated and perceived through the MVAE model, whose implementation we discussed in the previous section, and inserted in the RSA implemented model, as we will describe shortly.

Finally, we compare the base case results to the implementation with emotions to measure whether the introduction of emotion helped the disambiguation of the situation and in which form.

Scenario Our RSA implementation focuses on a specific scenario: an interaction between two human actors talking about the context they find themselves.

The base scenario can be described as follows:

Context: Two individuals, whom we will refer to as the Listener and the Speaker, are on a beautiful tropical beach on a sunny day. The Speaker emits an utterance.

Speaker's Action (Utterance): "Terrible!"

The "advanced" scenario in which we add the information resulting from the emotional expressiveness of the human face can be depicted as the base scenario in which the Speaker performs a second action:

Speaker's second Action (Facial expression): The speaker smiles.

Our model will then evaluate the situation from the point of view of the Listener: they will reason recursively, assuming that the internal structure of the Speaker's mind and the previous knowledge is similar to their own and, therefore, its state can be inferred from their behavior and the shared context to make assumptions about the Speaker's intention.

Base Case: In our theoretical introduction to RSA, we listed some elements that would shape the communicative act and enable the inference of intents between the two actors involved. Let us now model some of those elements and propose a possible

implementation:

From the definition of the scenario, we can easily define the **action** performed by the Speaker, that is to say, the uttering of the word "Terrible." A second action the Speaker performs in the "advanced" model is using facial expressions, a smile, to convey emotions.

Both Speaker and Listener share a set of concepts that categorizes together: the external inputs Z_{ext} coming from the external shared context ("Sunny day on a tropical beach"); the interoceptive sensations of the body, represented by the core affect $\mathcal{F} = (V, A)$ through the variables of valence V and arousal A; a series of pre-existing believes $\mathcal{B}$. This summarizes in the distribution $P(C|Z_{ext}, F, \mathcal{B}, \mathcal{L})$.

Defining the priors The utterance is a word selected from a subset of words $w \in \mathcal{L}$ where $\mathcal{L}$ represents the lexicon adopted. In our case, the adopted lexicon is a limited collection of just three words: "terrible," "OK," and "amazing". We associate a mental conceptualization of the external context to each word of our subset as perceived by the two actors. We call this mental representation the "state", defined as: $\mathcal{S} = \{s_1, s_2, s_3\} = \{terrible, OK, amazing\}$

We defined a conditional distribution that evaluates the probability of associating the current context ("Sunny day on a tropical beach") with each one of the states.

We define the state prior as $P(s) = Cat(s|\pi_1, \pi_2, \pi_3)$, were Cat is a categorical distribution and π_i represents the probability of assigning the state s_i.

State s	**"Terrible"**	**"OK"**	**"Amazing"**
Prior Prob. π	0.001	0.495	0.495

Table 1: State prior distribution for the current scenario.

Those values may change by changing the context or the previous knowledge and beliefs of the actors involved. The choice of those particular values is quantitatively arbitrary, and the intent behind it is to represent the imbalance in associating the word "terrible" with the context under consideration.

Each state is also associated with a probability distribution over the values of the core affect. This means that for each of the variables that compose the Core Affect $\mathcal{F}$, hence Variance V and Arousal A, we define a conditional distribution over the probability of getting low values of variance and high values of arousal for any given state. For simplicity, instead of modeling $\mathcal{F} = V \times A$ as a continuous space, we define the valence with only the binary alternative of positive and negative (+1, -1) and arousal with the "high" and "low" values:

$$v \in V = \{-1, 1\}$$
$$a \in A = \{low, high\}$$

Given a state s, it is possible now to define a conditional distribution of Valence and Arousal over s using Bernoulli distributions:

$$P_V(v|s) = Bern(v|\pi^V(s_i))$$
$$P_A(a|s) = Bern(a|\pi^A(s_i))$$

Those distributions may be expressed given the following values:

State s	**"Terrible"**	**"OK"**	**"Amazing"**
Low Valence Prob.	0.5	0.3	0.2
Positive Arousal Prob.	0.4	0.2	0.4

Table 2: Valence and Arousal prior distributions given the state

As we said before, those values represent a qualitative indication of the affective state given the "common" beliefs about each one of them. The valence indicates if the emotional impact of the state on the actor is positive or negative. Intuitively, a negative impression is more probable given the state "terrible" and less and less negative given the states of "OK" and "amazing". While the "terrible" and "amazing" states are assigned with a more pronounced probability for the "high" activation value (Arousal), the "OK" state represents a more mild emotional activation state and hence is assigned a lower probability.

In the base case where we do not consider the Speaker's facial expression, those prior values over the state for the Core Affect are shared between the Speaker and Listener since they belong to the shared beliefs. We will explore how to model the input from the Speaker's facial expression in the following section, where we will also compare the results of the two variations of the model.

4.2.1 Modeling the Speaker

Following, from the point of view of the Speaker, we define the communication goal it wants to achieve. It represents a particular meaning the Speaker desires to communicate to the Listener. To model the goal g, first, we have to define the concept

of meaning: The meaning depends on a set of current beliefs $\mathcal{B}$, and we can denote the meaning space as $\mathcal{M} = \mathcal{L} \times \mathcal{F}$. The goal g can be seen as a mapping from the entire meaning space $\mathcal{M}$ to one of its subset $\mathcal{M}_X$, according to the Speaker's desire: $g : \mathcal{M} \mapsto \mathcal{M}_X$

In our implementation, the full meaning space $\mathcal{M}$ is represented by the set of states $\mathcal{S}$, and the core affect space $\mathcal{F}$. The goal is, therefore, a subset of the state and affects variables, Valence and Arousal. We use a categorical distribution with uniform sampling between the possible outcomes to define the prior distribution on the goals:

$$g_i \sim Cat(g_i|\pi_i^G, i = s, v, g) \tag{12}$$

with π_i^G all equals.

The Speaker must also choose an utterance to communicate. For simplicity, since we already proposed mapping between words and the concepts expressed by the states we defined, we select the utterance by selecting one of the states' words. Even in this case, we use a categorical distribution over the states $\mathcal{S}$, with an equal probability between all the options:

$$u_i \sim Cat(u_i|\pi_i^U, i = s, v, g) \tag{13}$$

with π_i^U all equals.

4.2.2 Literal Listener

Every recursion must have a base case in order to terminate the computation, and the RSA model makes no exception. We have to define a basis for interpreting the Speaker's utterance, a procedure to interpret what is said straightforwardly, before building up the interpretation of eventual alternative meanings. We call this base case the Literal Listener (L_0).

L_0 's work is to interpret the Speaker's utterance literally, regardless of any interpretation about further aspects: the Literal Listener takes into account the Speaker's utterance and evaluates it according to its current state.

We can define a literal interpretation function for the utterance that given a state returns True if the state and the utterance coincide, hence $[[u]] : S \to Bool = 0, 1$.

$$[[u]](s) = \delta_{u=s} \tag{14}$$

The Literal Listener L_0, given an utterance u, defines the posterior distribution:

$$p_{L_0}(s, v, a|u, g) = p_V(v|s)p_A(a|s)p_S(s)\delta_{u=s} \tag{15}$$

The inference is performed through the repeated sampling from the prior distribution we already defined: We start sampling a state through a categorical distribution with probabilities π_i assigned as in table Table 1.

$$s' \sim P_S(s)$$

Then given s', we also sample values of Valence and Arousal through Bernoulli distributions, whose probability values are defined as in table Table 2.

$$v' \sim P_V(s')$$

$$a' \sim P_A(s')$$

Thanks to the posterior described in Equation 15, the Literal Listener can compute:

$$P_{L_0}(m_X|u) = \sum_{s',v',a'} \delta_{m_x=g(s',v',a')} P_{L_0}(s', v', a'|u, g) \tag{16}$$

This allows the Literal Listener to define the subset of desired meaning m_X given an utterance u, and at the same time, allows the Speaker to select the utterance u according to the desired meaning they want to convey.

4.2.3 The Speaker

The Speaker reasons recursively by simulating a Literal Listener and performing the inference of the optimal utterance to emit in order to communicate the desired meaning:

$$P_{S_1}(u|s, v, a, g) = P_{L_0}(s, v, a|u, g)P(u) \tag{17}$$

Where the Speaker S_1 selects the utterance through a softmax rule in the form: $P_{S_1}(u|s, v, a, g) \propto e^{\alpha U_1(u|s,v,a,g)}$ where α represents a parameter that determines the rationality of the speaker's choice. The higher the value of α, the more the Speaker's choice would approximate a utility maximization function.

4.2.4 Pragmatic Listener

The Pragmatic Listener, in turn, reasons recursively about the Speaker. It is worth noting that both the Speaker and the Literal Listener are models embedded inside the Pragmatic Listener's "mind". They are simulations implemented by the Pragmatic

Listener to understand the utterance expressed by the "real" actor it interacts with, assuming that both share a similar set of beliefs and a similar reasoning mind. Keeping that in mind, we follow illustrating the first step of the recursion where we describe the functioning of the Pragmatic Listener: The Pragmatic Listener evaluates the utterance emitted by the Speaker according to its prior beliefs through the repeated sampling of s, v, a, and g. Such action updates its beliefs by conditioning the resulting goal with the utterance u to find the set of s, v, and a that most likely represent the Speaker's internal state, hence the desired meaning.

$$P_{L_1}(s, v, a|u) = P_S(s)P_V(v|s)P_A(a|s)\sum_{g'}(u|s, v, a, g')P(g'|s, v, a) \tag{18}$$

4.2.5 Introducing emotions

In the base scenario, the only action performed by the Speaker was the utterance of the sentence. The Listener's interpretation of the internal emotional state, and therefore, the goal of the Speaker, can be based only upon the prior beliefs about the context and affective state. Now we explore how the emotions' expression influences the interpretation and eases the resolution of a communicative act. To do that, we model the "advanced" scenario where the Speaker performs a second action: The Speaker now smiles at the Listener while uttering the sentence "Terrible."

The Speaker's smile is possible through the activation of his facial muscles. People have an intuitive interpretation of the meaning behind each configuration of facial muscle activation, that is, facial expressions. We are intuitively able to map facial expressions with the emotional phenomena that originate those expressions, and that exactly is what we want our Pragmatic Listener to be able to do. Moreover, we want to provide a system able to perform the opposite, generating facial expressions starting from the emotional state. We developed those models through Multivariate Variational Autoencoders, and we can now use them to perform the Speaker's action and the Listener's interpretation.

Starting from a selected categorical emotion, we can use our models to encode such emotion into its latent form as a vector by using our $P(z|emotioncategory)$ encoder. Then we can use our $P(actionunit|z)$ decoder to generate the relative facial expression. This process represents the action performed by the Speaker. The Listener will perceive the facial expression and use the $P(z|actionunit)$ encoder to encode the facial expression into its latent form and the $P(categoricalemotion|z)$ decoder to extrapolate the resulting categorical emotion beneath it.

Since the actors' affective state is represented through the core affect's variable Valence and Arousal, we need a mapping from emotion category to VA representation. As discussed in chapter 2, the representations of emotions in computer science is an

explored field, and options to "translate" from one representation form to another are available. We can therefore use a map function that starts from categorical emotion sample values of Valence and Arousal.

This process can be represented by the following distributions:

$$P_{V_F}(v_F|AU) = Bern(v_F|\pi^V(AU))$$

$$P_{A_F}(a_F|AU) = Bern(a_F|\pi^A(AU))$$

Where AU represents a set of Action Units, we utilize for the interpretation of the facial expression. Those values of Valence and Arousal represent an additional source of affective information the Pragmatic Listener can utilize together with her set of prior beliefs. The Listener updates its beliefs about the affective state of the Speaker and performs a more accurate inference of the meaning the Speaker is trying to convey.

4.3 Results analysis

Let us now discuss the result of the simulations of the implemented models. First, we implemented the basic model in which the Pragmatic Listener infers the affective state of the Speaker only in function of prior knowledge given the context and the spoken utterance.

We performed multiple simulations in order to explore how the Listener interprets different Speaker's utterances. The constants for each simulation were the context and the corresponding prior knowledge. For the basic model, where no further information about the Speaker's affective stage was given, we performed three simulations, one for each possible utterance: "Terrible", "OK", and "Amazing".

The plots we obtained represent the resulting probability distribution of the inference. The x-axis reports all the possible configurations of the intended goal (the message's significance) that the Speaker wanted to convey through the communicative act; the y-axis represents the probability for each goal to denote the correct interpretation.

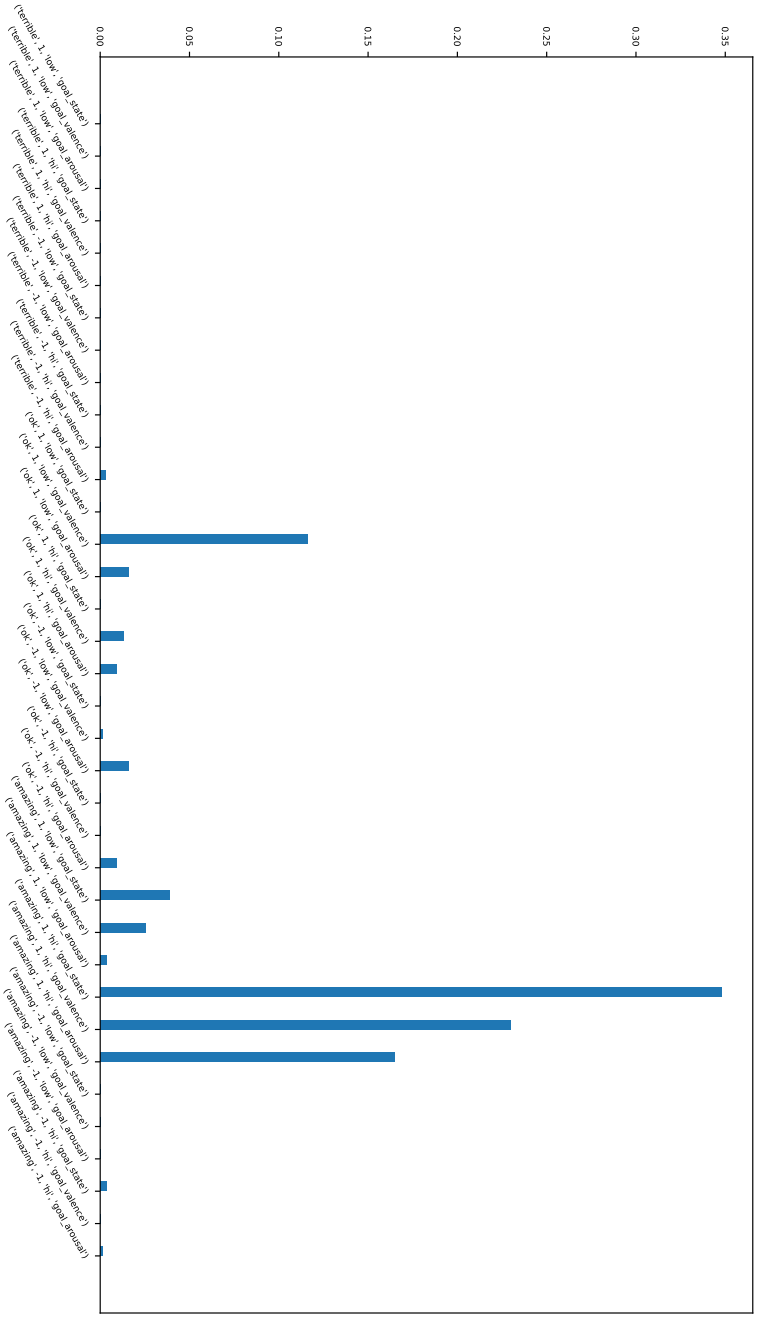

Figure 5: Posterior probability over the states for the base model given the utterance "Amazing".

In the case of the utterance "Amazing" (Figure 5), we notice how the most probable interpretations all involve a positive valence and an active state of arousal, regardless of the goal, with the intended meaning of the utterance coherent on what is spoken. Therefore, whichever aspect the Speaker is referring to (the state of the world, its arousal or valence state), he is in a positive, excited state, which is consistent with his utterance. In this case, the statement is coherent with the context; therefore, it does not represent a case involving irony, and the model correctly depicts it.

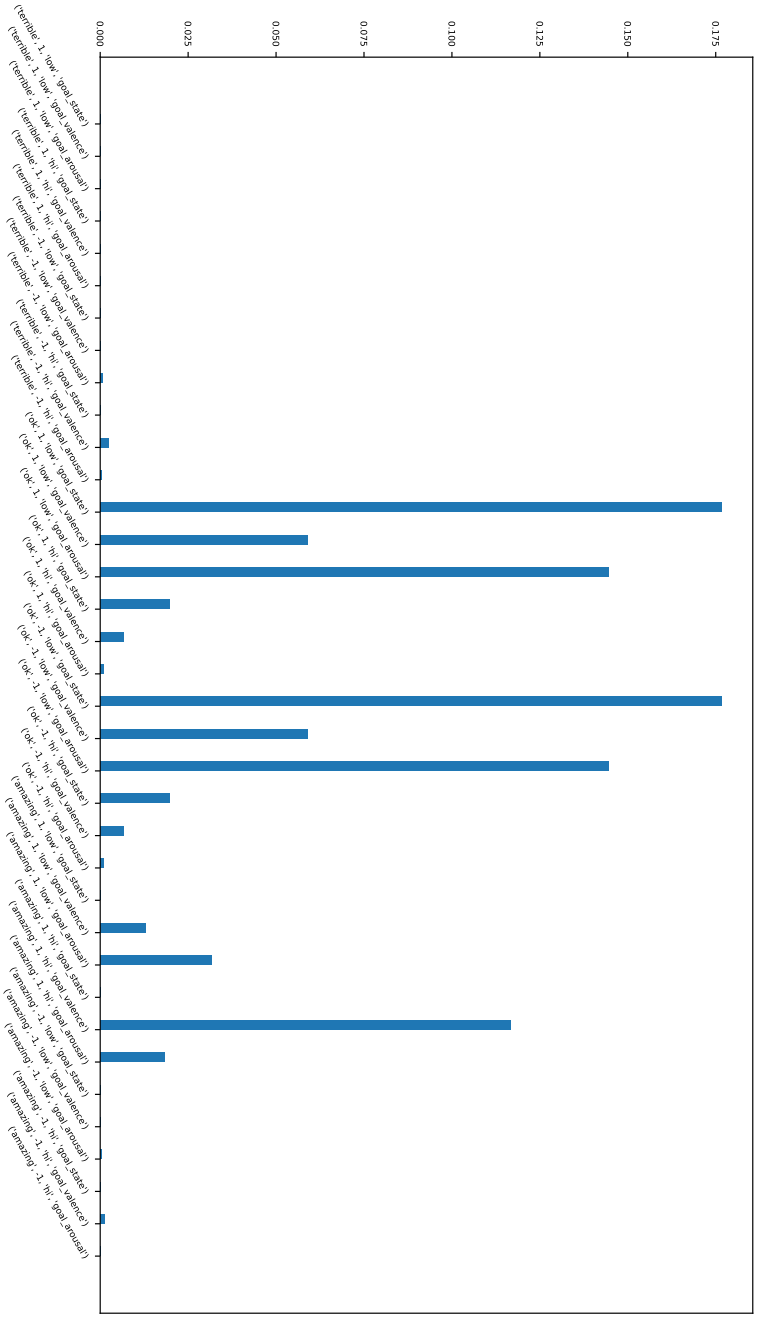
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
('terrible', -1, 'low', 'goal_state')
('terrible', -1, 'low', 'goal_valence')
('terrible', -1, 'low', 'goal_arousal')
('terrible', -1, 'hi', 'goal_state')
('terrible', -1, 'hi', 'goal_valence')
('terrible', -1, 'hi', 'goal_arousal')
('terrible', -1, 'low', 'goal_state')
('terrible', -1, 'low', 'goal_valence')
('terrible', -1, 'low', 'goal_arousal')
('terrible', -1, 'hi', 'goal_state')
('terrible', -1, 'hi', 'goal_valence')
('terrible', -1, 'hi', 'goal_arousal')
('ok', -1, 'low', 'goal_state')
('ok', -1, 'low', 'goal_valence')
('ok', -1, 'low', 'goal_arousal')
('ok', -1, 'hi', 'goal_state')
('ok', -1, 'hi', 'goal_valence')
('ok', -1, 'hi', 'goal_arousal')
('ok', -1, 'low', 'goal_state')
('ok', -1, 'low', 'goal_valence')
('ok', -1, 'low', 'goal_arousal')
('ok', -1, 'hi', 'goal_state')
('ok', -1, 'hi', 'goal_valence')
('ok', -1, 'hi', 'goal_arousal')
('amazing', -1, 'low', 'goal_state')
('amazing', -1, 'low', 'goal_valence')
('amazing', -1, 'low', 'goal_arousal')
('amazing', -1, 'hi', 'goal_state')
('amazing', -1, 'hi', 'goal_valence')
('amazing', -1, 'hi', 'goal_arousal')
('amazing', -1, 'low', 'goal_state')
('amazing', -1, 'low', 'goal_valence')
('amazing', -1, 'low', 'goal_arousal')
('amazing', -1, 'hi', 'goal_state')
('amazing', -1, 'hi', 'goal_valence')
('amazing', -1, 'hi', 'goal_arousal')

Figure 6: Posterior probability over the states for the base model given the utterance "OK".

In the case of the utterance "OK" (Figure 6), we see how the model prefers interpretations coherent with the spoken sentence, with a low arousal value. The results are similar for the positive or negative interpretation of the valence. The low level of affective activation is easily comprehensible: the utterance "OK" is associated with low activation values. The uncertainty regarding the valence can also be associated with the low activation value expressed by the utterance since, given the context, extreme stances are more likely, and the selection of a more neutral one could cause such indeterminacy.

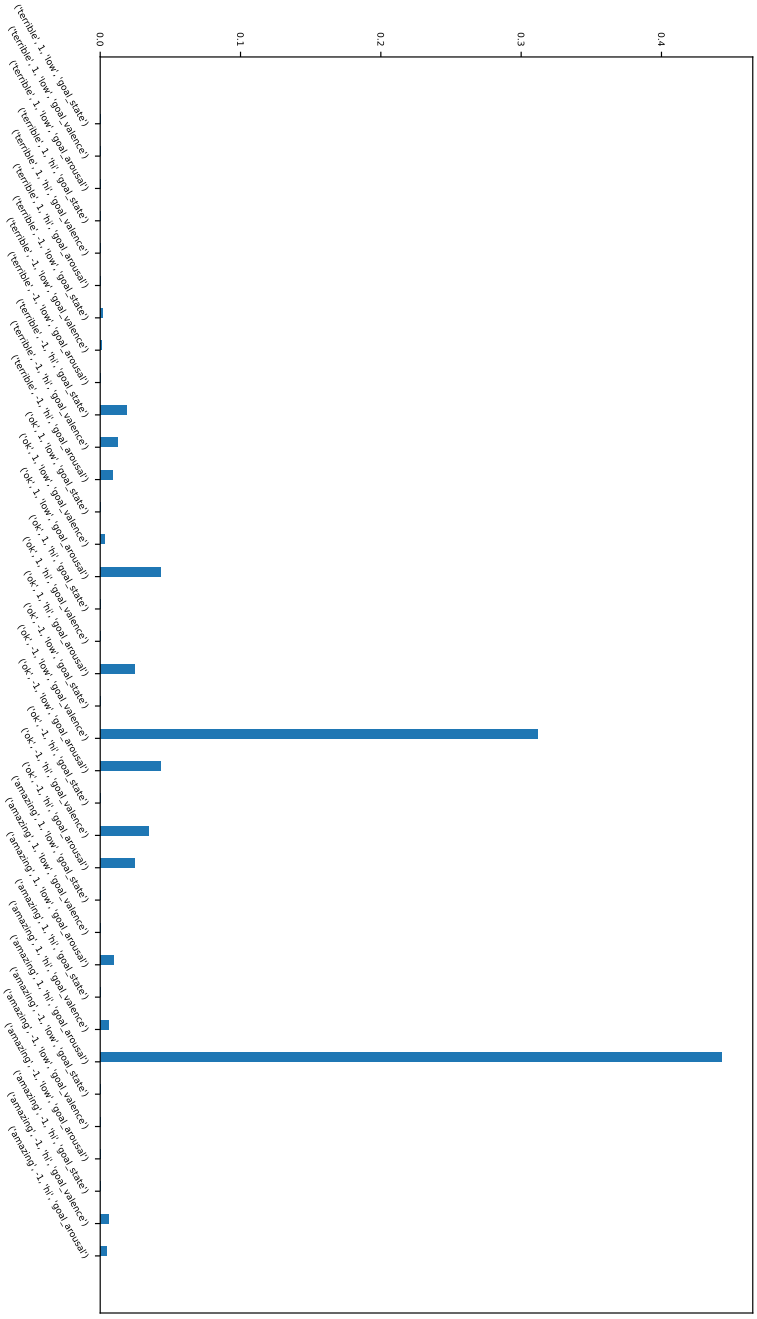

Figure 7: Posterior probability over the states for the base model given the utterance "Terrible".

Finally, for the basic model, we interpret the result in the case of the utterance "Terrible". This is the case where the expectations of the Listener should be disappointed, outlining a case of irony. By looking at the plot of Figure 7, we can clearly see how the most likely interpretation is represented by the case where the intended message is "Amazing", arousal is high, and valence is positive. This configuration directly contrasts with the uttered word and therefore conforms to the ironic interpretation of the communicative act. We also notice that a second (minor) peak represents a state of low activation, negative valence, and with "OK" as the intended message. This second configuration is coherent with the uttered word and in contrast with the ironic interpretation of the sentence. Ultimately this first model can spot irony but retains a certain degree of uncertainty. It is nonetheless a good starting point for further improvements.

Let us now analyze the advanced cases in which the Speaker uses the facial expression to communicate its affective state to the Listener. In this case, the possible simulations we could perform were many more since, for each utterance ("Amazing", "OK", "Terrible"), we had eight possible categorical emotions for a total of 24 possible simulations. We will now focus only on the more representative case that best represents the ironic statement: The scenario we designed our simulations around is where the Speaker utters the word "Terrible" while smiling. So we select "Terrible" as an utterance and a categorical emotion coherent with a smiling face, such as "Happiness".

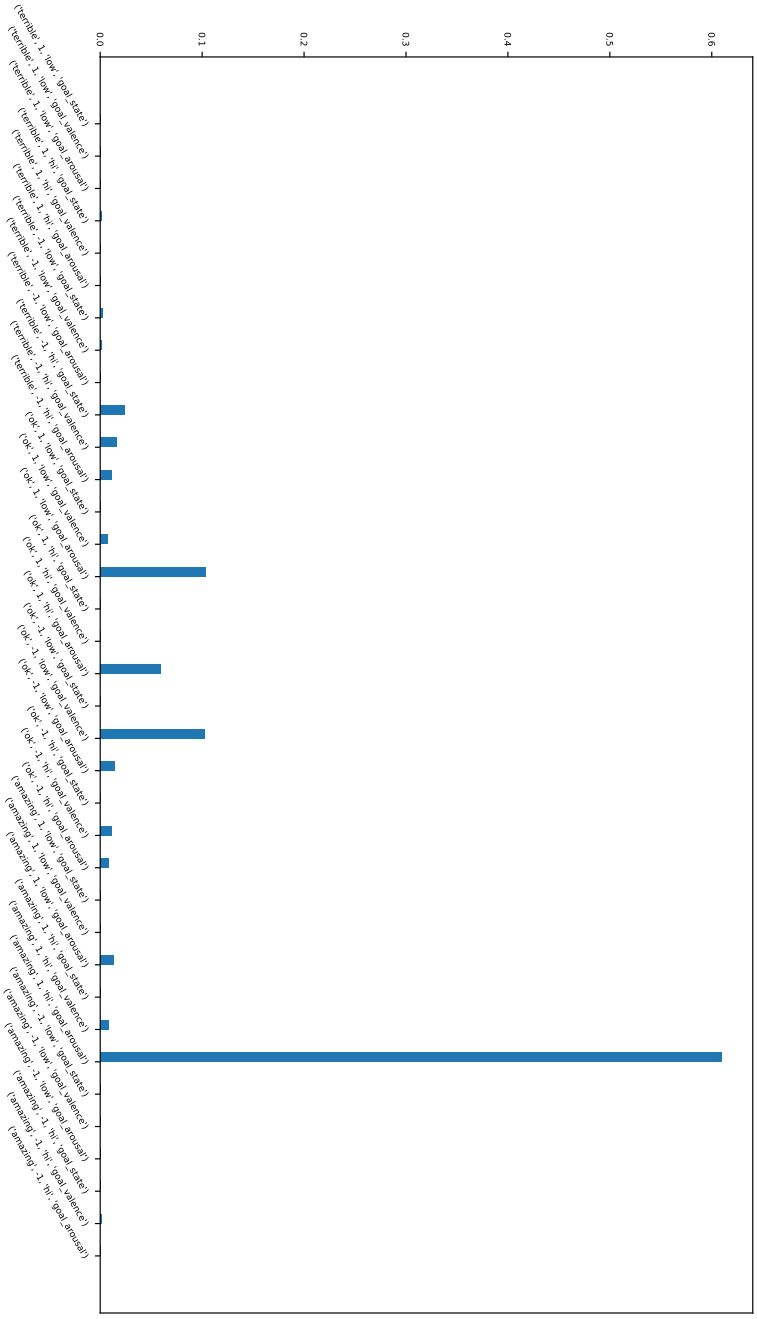

Figure 8: Posterior probability over the states given the utterance "Amazing" and the "happy" facial expression

From the plot of the (Figure 8) we can clearly see how the most probable interpretation of the utterance and the smiley face is by far the one that corresponds to a positive valence, high level of arousal, and the desired meaning diametrically opposed to the uttered word. It is undoubtedly how the introduction of emotions through facial expression helps the model to disambiguate ironic statements. We have a substantial and unequivocal improvement in interpreting the communicative act beyond spoken words.

Chapter 5

Conclusions

This thesis aimed to study the influence of emotion in spoken language and how they help interpret non-trivial forms of speech that require the use of implicature to reach a meaning beyond what the words directly say.

We approached our task by introducing the Rational Speech Act framework, which allowed us to model and implement pragmatic reasoning through an agent-based simulation of a communication act between two actors: a Listener and a Speaker.

Pragmatic reasoning means that the interpretation of a sentence does not only depends on the sentence itself but considers other elements such as context, prior knowledge, and the emotions conveyed through the communicative act.

RSA inherits from Grice's theories some of its principles, such as its recurrent structure and the cooperative structure for pragmatic communication. RSA is also inspired by the Conceptual Act Theory, where emotions are conceived as ad-hoc abstract categories and, therefore, deeply depend on the individual's affective state and its interpretation considering the current situation and the previous experiences. RSA accomplished take into account and implemented this theoretical approach into its conceptual-based structure through a Bayesian framework, where each concept is modeled through latent variables.

We then introduced the first model implementation of the RSA framework in which we desired to model an instance of a communication act involving irony. Irony is a form of speech in which the intended meaning of the uttered words is opposite to their actual intended purpose and represents perfectly a case where pragmatic reasoning could be used. We model an ironic statement by building a scenario in which an actor, the Speaker, utters a sentence in direct and obvious contrast with the actual context in which the agents are situated, and a second agent, the Listener, has to infer the real meaning behind the uttered sentence.

We build a second RSA model representing the same scenario as the previous one. Still, in this second case, the Speaker, as well as uttering a sentence as in the previous

scenario, communicates his affective state through facial expressions by smiling.

To model this second action performed by the Speaker, we utilize a generative model called Multimodal Variational Autoencoder. This model allows us to represent emotional concepts through a latent space from which we sample new instances of categorical emotions and images representing the corresponding facial expressions.

We trained and applied our MVAE module to the RSA model in order to introduce new affective information to the Listener's interpretation. To do so, we mapped the resulting categorical emotion to the corresponding core-affect values and used those values as affective cues to the Speaker's emotional state toward the depicted current situation.

5.1 Results

Comparing the new implementation to the initial simulation model, we obtained excellent results in the Listener's ability to understand the ironic statement thanks to the introduction of the emotional cues furnished by the Speaker's facial expression. The model yield the correct, ironic interpretation with a probability of 0.6 against the probability of 0.4 resulting from the simpler case with no facial expression cues. Additionally to that, the base simulation's result also yields a second interpretation with a significant probability of 0.3 that does not correctly represent the ironic context depicted; while for the case of the updated model, the most relevant interpretations without considering, second to the correct one, do not exceed the value of 0.1.

To briefly summarize the achieved results o this thesis:

- We introduced an implementation of the RSA framework that coherently depicted an ironic context to perform our simulations. This model serves as the base case to measure whether the introduction of affective information into the communicative act will help improve the interpretation of the conveyed message.
- We introduced the Multimodal Variational Autoencoder model to represent the concepts of emotions through different modalities, such as categorical emotion and facial expressions.
- During our model training, we addressed the problem of merging the encoding of different representation modalities, especially in the case of high differences in the number of features between each modality. Those problems brought us to a series of solutions to overcome or circumvent our issues: We adopted more advanced variants of the MVAE model, such as the $\beta-$VAE and Info-VAE models (obviously adapted to the multimodal context). We utilized action units instead of raw images as data points for our model.

- Finally, we applied our trained MVAE model to our initial RSA implementation for irony in order to analyze whether and in which measure the introduction of emotions in the communicative act allows for a better interpretation of non-trivial communicative context: our results confirm our initial statement, we can observe an improvement of the capability of our model to discern the ironic statement and correctly interpret it.

5.2 Potential improvements and future developments

Additional modalities to extract affective cues could be added to the model to improve the interpretation of utterances that requires implicature, such as the case of irony. Thanks to the modular nature of the probabilistic programming framework, it is possible to add new emotional sources to improve the model other than facial expression. For example, the study of prosody could furnish a new source of affective cues to integrate into the RSA model. But before adding new modalities to the "equation", new effort could be invested in improving the performance of the current MVAE models since integrating different input modalities has been a sore spot in the development of this project. The high number of hyperparameters and the number and differences in features slowed the development process. Up to this work's current state, the use of raw images as input in the MVAE model has been unsuccessful due to the low quality of the images generated. To circumvent those issues, we proposed extracting action units as a preliminary step to the model training. Such a fix has been demonstrated successful. Also, trying neural network structures and new ways to balance the input modalities could also provide better results to the overall model.

Appendix A

Introduction to Probabilistic Programming

Probabilistic programming is a paradigm that uses stochastic elements in computer programs to alter their functioning; the execution flow of a probabilistic program is not deterministic and results in a powerful tool to model real-world uncertainty and phenomena with a high degree of variability. Probabilistic programming allows the high-level implementation of statistical models while detaching the aspects related to the inference of the underlying probability distributions.

In their paper [18], Noah D. Goodman et al. show how probabilistic computing can model psychological-grounded theories directly into flexible and modular programs that can exploit the strengths of deep learning. Probabilistic programming could translate psychological theories of emotion into tractable computation models; once the psychological theory is translated into a statistical model, we can observe real-world data to infer the model's distributions. In this way, we can shape the concept of emotions at different levels of abstraction and deal with the uncertainty, randomness, and incomplete knowledge of this and similar phenomena while, at the same time, using real-world data to support the theoretical work.

Probabilistic programming has some advantages to the ad-hoc implementation of theory-based models: Theory-driven approaches tend to be hand-tuned to specific theories and contexts, which affects their flexibility and adaptability to a broader use; They usually do not specify/explain the steps between naturalistic data and the representation of the concepts they model (from the pixels that compose the image of a smiley person to the concept of the emotion extracted); While, on the other hand, data-driven approaches excel in the perception of different phenomena (e.g., recognition/classification of facial expression), they show some limits when we try to bound them to a more high-level theory.

By representing real-world phenomena with probability distributions, it is possible to take samples from them. Probabilistic programming is often coupled with the concept of generative models. A generative model can generate new data instances by sampling from the distributions that shape the model. Through a generative model, we can, for example, not only classify images based on their content but generate new images. We can evaluate events we perceive by implementing generative models that try to recreate the observed phenomena in their internal state and then infer their causes. The better the model can approximate reality, the more we can affirm that we were able to grasp the concepts and functioning behind those phenomena. These are some of the significant advantages over "classical" pattern-recognition techniques: generative models based on probabilistic programming help couple the theoretical knowledge with real-world samples by building models that encapsulate the concepts behind the data and the causal relations between them in a computable form.

The RSA framework and the MVAE model presented in this work exploit those principles: RSA uses probabilistic programming to model pragmatic reasoning. Through the MVAE model, we define the concept of emotion using a latent space from which we can sample either a categorical representation of the emotions (happiness, sadness, anger) or a visual representation of the emotion in the form of images.

Appendix B

Technologies and External Tools

B.1 Technologies used

For implementing the models described in Chapter 4, we utilized python version 3.10.4 as programming language, together with Jupyter Notebook as the primary developing tool.

To enable GPU computing, we used a machine mounting a Tesla V100S with 32Gb video memory and CUDA Version: 11.6 for parallel processing.

B.2 External Tools

In some cases, we used external tools to extract interesting features from the dataset files. In particular, we experimented with Action Units as an alternative to raw images as an input source for our MVAE model. We used OpenFace to extract action units from image files and OpenFACS to produce a 3D model to express the action units generated by our model.

B.2.1 OpenFace

OpenFace [3] is a computer vision tool that allows the extraction of different features from image and video sources related to facial behavior analysis. It allows the extraction of several different features from input files, such as facial landmarks, head pose, facial action units, and aye gaze estimations. For our work, we utilized only the facial action units extraction tool, which generated a set of 18 action units (namely the 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28 and 45). To be noted, action unit 28 was not provided by the tool, so the total number of features retrieved was 17. Action units are represented in two ways: by presence, with a boolean value, or by the intensity in a 5-point scale where 1 represents the minimum intensity and

5 represents the maximum intensity. In this second case, the absence of the action unit is reported with the value 0.

B.2.2 OpenFACS

OpenFACS [4] is the tool we used to visualize the action units generated from the MVAE model. From the public repository page:

> OpenFACS is an open source FACS-based 3D face animation system. OpenFACS is software that simulates realistic facial expressions by manipulating specific action units defined in the Facial Action Coding System. OpenFACS has been developed with an API to generate real-time dynamic facial expressions for a three-dimensional character. It can be easily embedded in existing systems without prior computer graphics experience.

OpenFACS utilizes the same set of 18 action units also utilized by OpenFace. The exceptional action unit 28 not provided by the OpenFace tool is ignored, forcing a constant value of 0 for each example.

B.3 Modalities merging

One crucial aspect when dealing with Multimodal Variational Autoencoders is the choice of the merging strategy, namely how multiple encoders (one for each modality) can generate one single latent vector representing the latent concept expressed by the input data.

In our tests, we tried three techniques: Product of Experts, Mixture of Experts, and merging through Neural Network.

B.3.1 Product Of Experts

Product of Experts[19] is one technique that allows producing a target probability distribution by multiplying several starting distributions, namely the experts, and then re-normalizing the output to sum to one.

PoE is characterized by the "veto" power of the experts, which means that if a single expert assigns low values to a particular value, then that value will have a low probability also in the resulting distribution.

$$p(y|x) = \frac{1}{Z} \prod_i p_i(y|x) \tag{19}$$

In our case, where the probability distributions are Gaussians in the form:

$$p_i(y|x) = N(\mu_i(x), \Sigma_i(x)) \tag{20}$$

the resulting PoE distribution is still a Gaussian, whose mean (μ) and covariance (Σ) are defined as follows:

$$\mu(x) = (\sum_i \mu_i(x) T_i(x))(\sum_i T_i(x))^{-1} \tag{21}$$

$$\Sigma(x) = (\sum_i T_i(x))^{-1} \tag{22}$$

with $T_i(x) = \Sigma_i^{-1}(x)$

B.3.2 Mixture of Experts

The Mixture of Expert (or MoE) is the second model we consider for the task of "conciliating" the results of the encoding of different modalities.

MoE is an ensemble technique, mainly adopted with neural networks and predictive modeling. It allows us to divide the main task into many sub-tasks and associate each with an "expert," a neural network trained for that specific task. The experts' predictions are finally combined to generate the final result.

While the Product of Expert technique is characterized by its "veto" policy, the Mixture of Experts adopts the opposite approach: a value can yield a high probability in the final distribution even if just one expert assigns it a high probability.

It follows the standard equation for the mixture of experts:

$$p(y|x_i) = \sum_i \alpha_i p_i(y|x_i) \tag{23}$$

$\alpha(x)$ is called mixture coefficients or gating functions; in our case, it is fixed to one. The mixture of experts combines the different components by addition and then normalizes the result according to the number of different modalities.

B.3.3 Neural Network Merging

The last merging technique we tried was the merging through the neural network. In this case, we adapt the encoder of the modalities we decide to merge so that the last layer is composed of a number n_i of output neurons. Each encoder will produce an output vector of size n_i. We use an additional dense, fully connected neural network that takes, as input, the concatenated vectors resulting from the encoders. This last neural network, finally, returns as output the parameters (μ and σ^2) of the normal

distribution from which we will then sample the latent vector.

Since none of the listed options outperformed the others in the field, we considered the merging technique a hyperparameter to be evaluated in the tuning phase of the models.

Appendix C

MVAE model: implementation details

In this appendix, we illustrate in detail the implementation of the MVAE models described in this thesis: we first report and describe the elements that compose the MVAE model, hence the neural networks that form the encoder and the decoder, the data used for the training for the different modalities, the hyperparameters that define the model. We then illustrate the training process and the evaluation strategies of the trained models. Finally, we illustrate the performances of the model implemented.

C.1 Model architecture

MVAE model comprises one encoder and one decoder for each data modality. In our case, we built two MVAE models, and each of them encodes two modalities:

1. MVAE model for categorical emotions and raw images
2. MVAE model for categorical emotions and action units

In the first MVAE model, we pair categorical emotions and raw images through the latent space; In the second one, we substitute the pictures with the Action Units extracted from raw images.

Raw images model We use deep convolutional neural networks to encode and decode features from the raw images. Those networks are inspired by the structure of those used in DCGAN models.

The encoder parametrizes the distribution $q(z|x)$, where x represents the input image, and z is the latent vector. The encoder takes as input an image shaped as a 3x64x64 tensor and returns as output two values representing the mean ad variance of the corresponding normal distribution.

The network is formed of three "convolutional blocks," each formed by one **convolutional layer**, **batch normalization layer**, and a **Leaky ReLU** activation function.

The kernel's size is three by three pixels wide, with a padding value of one pixel to maintain the tensor's shape. **The filter's number** for each layer is specified as a hyperparameter to ease the tuning process. However, each layer increases the number of filters by a multiplicative factor of 2 to the previous layer.

After each "block", a **max pool layer** is inserted between the layers to reduce the width and height size of the tensor while the number of filters in each layer increases. A final convolutional layer closes the sequence of convolutional blocks.

After the sequence of "convolutional blocks," the neural network's structure forks into two sequential blocks of fully connected layers to define the output normal distribution's parameters (mean and variance).

The image decoder is used to parametrize the distribution $p(x|z)$; its shape is specular to the encoder's network: First, we have a sequential, fully connected layer that takes as input the latent vector z and whose output is a three-dimensional tensor ready to be processed by the sequence of convolutional blocks to generate an image x as output. The kernels' number doubles after each convolutional block in the encoder, halved in the decoder where the structure is reversed. Each max pool block is also replaced with an upsampling layer that increases the tensor's width and height by two.

Action Units model Compared to raw images, action units can be represented in a more compact form. It is sufficient for a vector of size 17 to represent a facial configuration that can express emotions instead of the 3-dimensional 3x64x64 tensor required for the raw image. Moreover, images require the use of convolutional networks, which are a specialized type of neural networks that exploits the spatial position of each value in relation to others, while when dealing with simple one-dimensional vectors, such as in the case of AUs, a much simpler dense neural network is more appropriate.

As in the previous case, the encoder parametrizes the distribution $q(z|x)$. Is input is the 17-dimension tensor representing the action units, and the output is the mean and variance parametrizing the normal distribution from which the latent vector is sampled.

The input is mapped from the input layer to the first hidden layer of size h, where h is a hyperparameter defined during the model tuning. After that, the network structure is forked into identical branches, one to learn the value of the mean and one to learn the value of the variance of the resulting normal distribution.

Each branch comprises two fully connected linear layers, followed by one batch normalization layer of size h and a leaky ReLU activation function. The final output is mapped to the size of the latent space z through a final fully connected linear layer.

The action units decoder is a simple, fully connected, four-layer deep, multilayer perceptron that maps the input latent vector z into the AU vector of size 17.

Categorical emotion encoder and decoder structure: Those two models share the same categorical emotion modality; hence, the encoder and decoder for such cases are shared. Encoder and decoder for categorical emotions share the exact structure of the AUs ones since the input shape, and output shape are similar. Only the vector size to represent emotion is smaller, which follows the reduced number of layers required for categorical emotions concerning AUs.

C.2 Model hyperparameters

The training of the Multimodal Variational Autoencoder required the tuning of several hyperparameters:

- Latent space dimension: The choice of the size of the latent space depends on the task approached and the number of features of the encoded data: autoencoders are usually applied for dimensionality reduction, mapping the input into a smaller latent space, but some applications may benefit from a larger latent space, for the input, such as autoencoders applied to image denoising. Since we do not fall in either the first or second categories of application described above, we have to decide the size of the latent space based on our necessities.

- Hidden layer dimension: This hyperparameter defines the size of the hidden layer of all the neural networks (encoders and decoders), hence the layers between the input layers and the output layers. The size of layers could affect the capability of the network to learn new features. By increasing and decreasing this value, we could also handle (within certain limits) the effects of overfitting and underfitting. Usually, the dimension of the hidden layer is decided when the general structure of the neural network itself is defined. In this case, we opted to allow this particular factor to be separated and handled independently to explore how the hidden layer's size influenced the overall results. The neural network structure is a complex topic, and the possible configurations are endless. For this reason, when dealing with image handling, we tried to limit our choice to a few models supported by the literature and known for their good performances in similar jobs.

- Number of filters: The number of filters (or kernels) is a parameter used only when dealing with images. In those cases, we adopted convolutional neural networks as encoders and decoders, which required the definition of filters or kernels. Since the filters' number's choice as hyperparameters for the model

definition is similar to the one for the hidden layer size, the motivations behind such are analogous.

- Alpha: parameter introduced with the modified ELBO loss function for implementing the Info-VAE model.
- Beta: parameter introduced to improve the training of variational autoencoders [20]. Essentially, we use a multiplicative factor to reduce or increase the influence of the Kullback–Leibler loss value. Citing the original paper, the researchers introduced an

  > "adjustable hyperparameter beta that balances latent channel capacity and independence constraints with reconstruction accuracy.

- Reconstruction weights: Multiplicative factors are also applied to the loss values forming the reconstruction loss. The reconstruction loss is calculated by summing the loss of each modality (face images/action units, emotion categories). By multiplying each value by a constant factor, it adapts the influence of each modalities' error, increasing or decreasing the importance of each of them in the computation.

Through initial training tests and evaluating similar models in the literature, we selected a reasonable pool of values for each hyperparameter to search for the best configuration:

Latent Space Dimension	[10, 256]
Hidden Layer Size	[16, 512]
Number of Filters	[16, 512]
Learning Rate	[1e-7, 1e-3]
Alpha	[1e-7, 10]
Beta	[1e-7, 10]
Reconstruction Weights	[1e-4, 1e+6]
Number of Epochs	[25, 100]
Batch Size	[8, 512]

Due to the number of hyperparameters and their possible values, a complete search of the optimum model is not feasible. We used automatic tools for hyperparameter tuning, such as Torch Ray Tune. The tool required the specification of an interval or a set of possible values to run the training and search for the best combination within a predefined maximum number of trials. However, such an option has not always been possible. Due to the size of the images, we used to train one of the versions of the variational autoencoder, training a specific configuration of parameters

could take more than one hour, also depending on the size of the neural network involved. In this case, we opted to search for a more limited number of possible configurations of the hyperparameters, trying to find the correct values for each of them by reasoning over the results of the previous attempts and filtering out the ones that gave the worst results. The limits of this approach are that possible good combinations of hyperparameters may not be explored or discarded due to the low performances of similar configurations. Nevertheless, there is no silver bullet for training neural networks, only best practices and founded reasoning within the scope of our possibilities.

C.3 The Dataset

As a dataset for our model training, we used RAVDESS [21]. RAVDESS contains 7356 files reporting 24 professional actors vocalizing two lexically-matched statements in a neutral North American accent. Each actor performs each statement multiple times, expressing different emotions. The emotions considered for this project are: calm, happy, sad, angry, fearful, surprise and disgust. The emotions presented broadly match the natural emotions described by Paul Ekman in his works. Each emotion is conducted through facial expressions and voice prosody, performed through speech and singing. Each actor performs at two intensity levels for each expression (normal and strong).

The produced files are available in three formats: audio-only, audio-video, and video-only. We used video-only files since our model relied on images. However, the theoretical base we applied could be adapted for other modalities, such as audio, to exploit multiple information sources.

C.4 Data preprocessing

We produced two data sets from the RAVDESS video files to train the MVAE model with different input data modalities.

The first dataset comprises 12 randomly chosen, evenly separated frames extracted from each video-only file and saved as a png file. Those raw images have been resized to 64 by 64 pixels and centered on the actor's face. Each image has been associated with a name that reports the actor's id, sex, and the represented categorical emotion. From each image, we produce a 3-dimensional tensor of shape 3x64x64, where the first dimension represents the RGB values of each pixel that composes the image. As the last step, we applied a normalization to the tensor to obtain the [0,1] interval of values. This tensor is then ready to be processed by the convolutional neural network that forms the encoder of the MVAE model. The resulting dataset comprises 9867

pairs of vectors and emotion categories. The emotion categories associated with each vector represent the ground truth label of the image and can assume an integer value between 0 and 7. Each value is associated with a discrete emotion such as 'neutral', 'calm', 'happy', 'sad', 'angry', 'fearful', 'disgust' and 'surprised'.

The second dataset is generated by submitting the frames of the first one to the OpenFace tool to extract the face action units from the images. We obtained a 17-dimension vector for each image composed of the extracted action units associated with a label representing the emotion category associated with the starting image. We proceed then to normalize the vectors in the [0,1] interval.

C.5 Model training and results

Each Multimodal Variational Autoencoder model is trained through supervised learning; the data set is composed of tuples of data entries. Each tuple contains different modalities encoded into the same latent space and represents the same concept; in our case, each tuple consists of an image/AUs and the categorical representation of that same emotion expressed by the image. The model should learn to associate the modality entries with each other and each modality with itself alone(such in the case of the normal autoencoder). To do this, we submit all possible combinations of inputs and ground truth.

C.6 Training Evaluation

Our model expects the ability to extract the concept of discrete emotions and recognize and generate images depicting one of such emotions. For this reason, we decided to evaluate the performance of the models according to two different tasks:

1. Given the image of a person's face expressing an emotion, correctly categorize such image according to the emotion expressed.

2. Given a categorical emotion, correctly generate a recognizable image depicting such emotion.

The evaluation of the first task is easy since each image in our data set is already associated with a categorical label, and we have to check if the generated label corresponds with the ground truth. Given the classification results, it is easy to generate a confusion matrix to address the result of the training.

The second task is more challenging since the face image generated has to be evaluated manually. We can assert that a good model should correctly classify the

images it generates. We can therefore exploit this principle to automatize the evaluation and obtain a similar metric for the classification task. In other words, we can let the model evaluate the same images it generates by choosing a categorical emotion label and using the Categories Encoder to generate the latent vector of such emotion. We then use the Images Decoder to generate a corresponding image. We submit the generated image to the Images Encoder to sample a new latent vector and obtain a categorical emotion label by submitting the latent vector to the Categorical Decoder. Finally, we can compare the starting categorical emotion with the generated label. Repeating this process multiple times for each emotion category gives us a good idea of how well the model can generate images starting from categorical values, similar to the classification task.

Although this strategy can give us interesting insight, it is not a solid metric to rely on. Due to the high dimensionality of the data, the model may generate and evaluate image details that are not relevant from the point of view of a human actor, learning to encode and decode features that do not hold meaning except for the model itself.

For this reason, we only use the above metrics to formulate a general idea of the model's performance. Final human control is required to evaluate whether the generated images met the standard of human judgment, both in terms of the general quality of the image and recognition of the emotional expression.

Images-Categories training The first model is trained according to the original idea of using images of real person's faces from the RAVDESS data set.

According to the evaluation strategies we previously defined, the best model is obtained using the following set of hyperparameters:

Latent Space Dimension	50
Hidden Layer Size	256
Number of Filters	128
Faces Reconstruction Weight	1
Categorical Reconstruction Weight	10000
Number of Epochs	50
Batch Size	32
Expert Type	Mixture of Expert
Learning Rate	5e-6
Alpha Weight	1
Beta Weight	1e-6

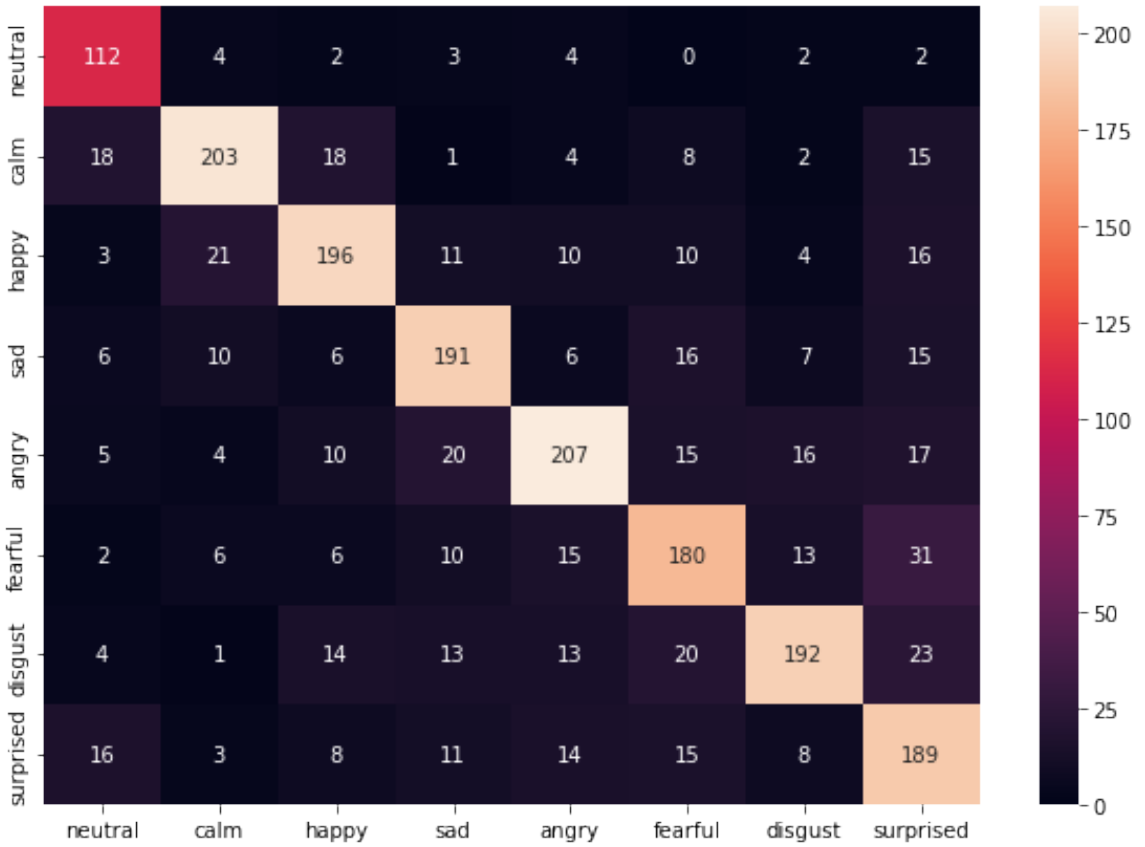

Figure 9: Confusion matrix for the Raw Images/Categorical Emotions model.

The model's performance in the classification task scored an accuracy of 0.73. The model's performance in the generation task is less impressive: the accuracy score for the generation is 0.55, but the images generated are confused and lack most of the human facial features. It is impossible to read human emotions. It is worth noting that the overall structure of the human face is present, but the quality of the generated image is poor and cannot be considered satisfactory. The model has been able to correctly classify more than half of the generated images over eight different categories, even though no human actor would probably be able to recognize anything on them. This result confirms our first intuition by demonstrating that our way of evaluating image generation performance is partially flawed.

Causes for this behavior may be due to the high dimensionality of the input images and the dimensionality differences between the two modalities; image vectors are formed of 64x64x3 values while categorical vectors are of size 8. Due to that, for each value specified by each emotion category, we have very few constrictions

imposed on the sampling of the new image. Image features other than emotions like age, sex, hair, and every other uncorrelated facial feature are free to be sampled during the generation step. Moreover, since the dataset is formed from the performance of twelve different actors with very different overall appearances, the features (unrelated to emotion) the model has been able to learn were not enough to correctly shape and represent the human face.

Action-Units training To overcome the previously cited issues, we propose a second model. This time instead of directly using the images of actors expressing emotions through facial expressions, we first try to capture those facial cues, called Action Units (AUs), and process those AUs as a modality for our multimodal variational autoencoder. Using AUs instead of raw images, we significantly reduce the number of features used to represent this modality that goes from 64x64x3 values to represent each image to a vector of 17 entries. Furthermore, by using external tools to isolate the emotional features (both in the extraction from the image and in the generation of new images), we can discard all the additional information that we believe may cause the issues in the image generation. OpenFACS uses a 3D model that allows us to modify 18 different Action Units and represent the resulting profile.

In this second case, due to the reduced number of features, the resulting model was much faster to train than the previous one. This improvement in the training time allowed us to use hyperparameter tuning tools that automatized the search for the perfect set of hyperparameters by training multiple model versions and saving the best results.

According to the evaluation strategies we previously defined, the best model is obtained using the following set of hyperparameters:

Latent Space Dimension	16
Hidden Layer Size	512
AU Reconstruction Weight	3950
Categorical Reconstruction Weight	0.01
Number of Epochs	50
Batch Size	256
Expert Type	Product of Expert
Learning Rate	0.001
Alpha Weight	0.1
Beta Weight	1e-7

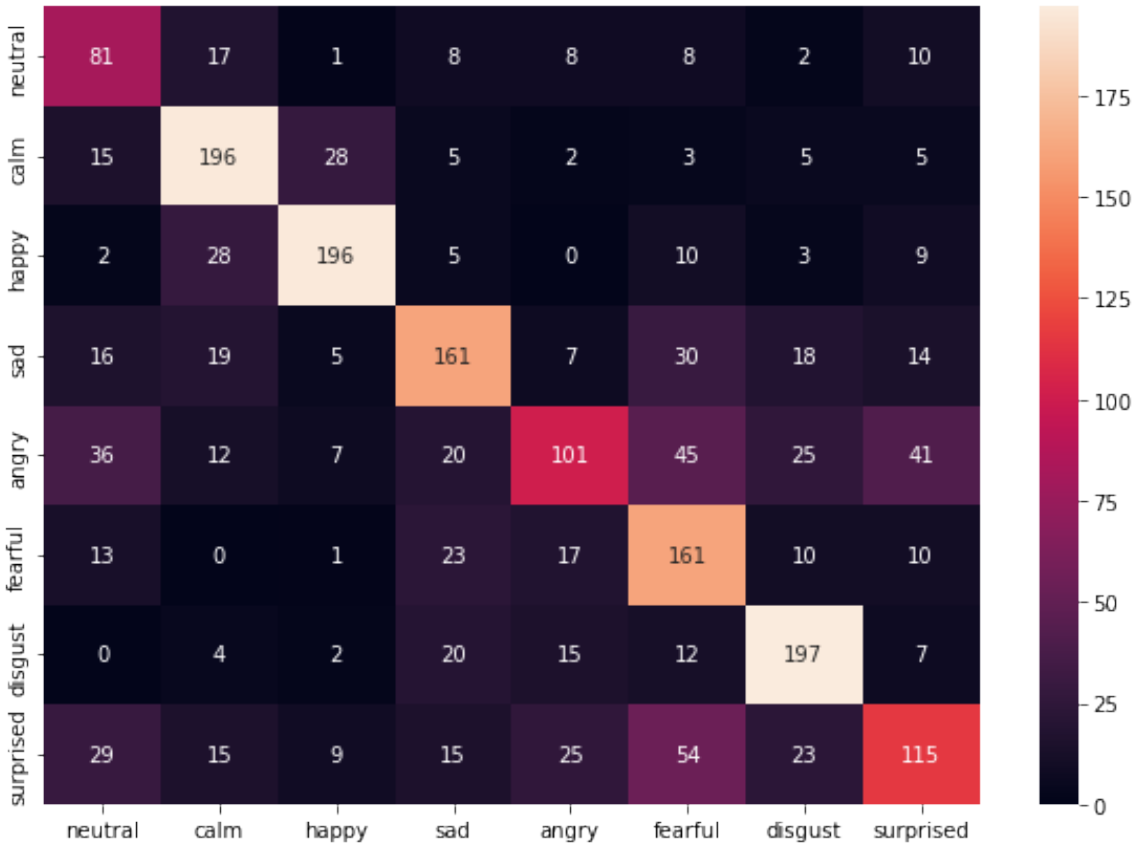

Figure 10: Confusion matrix for the Action Units/Categorical Emotions model.

In this second case, the generated images were much more recognizable, and the different emotions were identifiable. The classification accuracy score is 0.61, a less considerable performance with respect to the previous more complex model, and the reconstruction accuracy score is 0.8. The reconstruction score is very high with respect to the classification score; we can interpret this result as if the model generates faces with less variance compared to the input data. This result could partially confirm our initial suspicion of the issues for our first model. By reducing the number of possible facial features, selecting one categorical emotion influences a reduced (and more recognizable) number of possible representations of such category in the latent space and the AUs/Images space. The reduced performance on the classification task may be due to the additional processing of the data set and the introduction of additional noise: the dataset is formed by frames taken randomly from the performance of different actors. The facial expression may vary considerably during the act, as is

normal during a speech. Not all still images are "iconic" representations of the emotion depicted through the whole performance. It introduces an unavoidable noise in the data that may be enhanced during the extraction of the AUs performed through OpenFace (an external tool).

However, the model's performance and the quality (and recognizability) of the figures generated are sufficient to allow the use of our latest model for the scope of our project.

Bibliography

[1] Noah D Goodman and Michael C Frank. "Pragmatic language interpretation as probabilistic inference". In: *Trends in cognitive sciences* 20.11 (2016), pp. 818–829.

[2] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[3] Tadas Baltrusaitis et al. "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: May 2018, pp. 59–66. DOI: 10.1109/FG.2018.00019.

[4] Vittorio Cuculo and Alessandro D'Amelio. "OpenFACS: An Open Source FACS-Based 3D Face Animation System". In: *Image and Graphics*. Cham: Springer International Publishing, 2019, pp. 232–242. ISBN: 978-3-030-34110-7.

[5] Mary Helen Immordino-Yang and Antonio Damasio. "We feel, therefore we learn: The relevance of affective and social neuroscience to education". In: *Mind, brain, and education* 1.1 (2007), pp. 3–10.

[6] Antoine Bechara, Hanna Damasio, and Antonio R. Damasio. "Emotion, Decision Making and the Orbitofrontal Cortex". In: *Cerebral Cortex* 10.3 (Mar. 2000), pp. 295–307. ISSN: 1047-3211. DOI: 10.1093/cercor/10.3.295. eprint: https://academic.oup.com/cercor/article-pdf/10/3/295/9751042/100295.pdf. URL: https://doi.org/10.1093/cercor/10.3.295.

[7] Sean Spence. "Descartes' error: Emotion, reason and the human brain". In: *BMJ* 310.6988 (1995), p. 1213.

[8] Desmond C. Ong, Jamil Zaki, and Noah D. Goodman. "Affective cognition: Exploring lay theories of emotion". In: *Cognition* 143 (2015), pp. 141–162. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2015.06.010. URL: https://www.sciencedirect.com/science/article/pii/S0010027715300196.

[9] Paul Ekman and Wallace V Friesen. "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (1978).

[10] Lisa Feldman Barrett et al. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements". In: *Psychological science in the public interest* 20.1 (2019), pp. 1–68.

[11] James A Russell. "Core affect and the psychological construction of emotion." In: *Psychological review* 110.1 (2003), p. 145.

[12] James A Russell and Lisa Feldman Barrett. "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." In: *Journal of personality and social psychology* 76.5 (1999), p. 805.

[13] Katie Hoemann et al. "Developing an understanding of emotion categories: Lessons from objects". In: *Trends in Cognitive Sciences* 24.1 (2020), pp. 39–51.

[14] Lisa Feldman Barrett, Christine D Wilson-Mendenhall, and Lawrence W Barsalou. "The conceptual act theory: A roadmap." In: (2015).

[15] Justine T. Kao and Noah D. Goodman. "Let's talk (ironically) about the weather: Modeling verbal irony". In: *Cognitive Science* (2015).

[16] Geoffrey Hinton. "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Computation* 14 (2000), p. 2002.

[17] Shengjia Zhao, Jiaming Song, and Stefano Ermon. *InfoVAE: Information Maximizing Variational Autoencoders*. 2017. DOI: 10.48550/ARXIV.1706.02262. URL: https://arxiv.org/abs/1706.02262.

[18] Desmond C Ong et al. "Applying probabilistic programming to affective computing". In: *IEEE Transactions on Affective Computing* 12.2 (2019), pp. 306–317.

[19] Yanshuai Cao andDavid J. Fleet. "Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions". In: *CoRR* abs/1410.7827 (2014). arXiv: 1410.7827. URL: http://arxiv.org/abs/1410.7827.

[20] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[21] Steven R. Livingstone and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PLOS ONE* 13.5 (May 2018), pp. 1–35. DOI: 10.1371/journal.pone.0196391. URL: https://doi.org/10.1371/journal.pone.0196391.

[22] Katie Hoemann et al. "Developing an Understanding of Emotion Categories: Lessons from Objects". In: *Trends in Cognitive Sciences* 24.1 (2020), pp. 39–51. ISSN: 1364-6613.

[23] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. URL: https://doi.org/10.10802F01621459.2017.1285773.

[24] P. Grice. *Studies in the Way of Words*. ACLS Humanities E-Book. Harvard University Press, 1989. ISBN: 9780674852716. URL: https://books.google.it/books?id=QqtAbk-bs34C.