

An Inherently Explainable Model for Video Activity Interpretation

Sathyanarayanan N. Aakur, Fillipe DM de Souza, Sudeep Sarkar

Department of Computer Science and Engineering,
University of South Florida, Tampa

Abstract

The ability of artificial intelligence systems to offer explanations for its decisions is central to building user confidence and structuring smart human-machine interactions. Understanding the rationale behind such a system's output helps in making an informed action based on a model's prediction. In this paper, we introduce a novel framework integrating Grenander's pattern theory structures to produce inherently explainable, symbolic representations for video activity interpretation. These representations provide semantically coherent, rich interpretations of video activity using connected structures of detected (grounded) concepts, such as objects and actions, that are bound by semantics through background concepts not directly observed, i.e. contextualization cues. We use contextualization cues to establish semantic relationships among entities directly hypothesized from video signal, such as possible object and actions labels, and infer a deeper interpretation of events than what can be directly sensed. We demonstrate the viability of this idea on video data primarily from the cooking domain by introducing a dialog model that uses these interpretations as the source of knowledge to generate explanations grounded in both video data as well as semantic connections between concepts.

Introduction

Intelligent agents have evolved tremendously and have achieved significant milestones such as approaching human capabilities in some domains ((Kheradpisheh et al. 2016)). However, despite these performance gains, the model's ability to *explain* their decision appears to be constrained. Such ability to express the rationale behind its decision is vital when deploying models in an open, uncontrolled setting. For example, when taking vital decisions in high-risk areas like medical diagnosis (Caruana et al. 2015; Linder et al. 2014) and surveillance (Mahadevan et al. 2010; Junior, Musse, and Jung 2010) to name a few, the level of interaction between the human and a model is of high importance. It has also been established that a model with higher explainability is more likely to be trusted (Ribeiro, Singh, and Guestrin 2016) than a model with limited or no explainability.

Explainable models have been explored to some extent in literature. Spanning a variety of application domains such as

medical diagnosis (Shortliffe and Buchanan 1975), activity simulations such as those in the military (Core et al. 2006; Lane et al. 2005) and robotics (Lomas et al. 2012), these approaches have advocated models that are able to explain the approach undertaken to arrive at decisions but were not able to *justify* their decision to the user. There also have been *model-agnostic* approaches such as ((Baehrens et al. 2010; Ribeiro, Singh, and Guestrin 2016)) that attempt to explain the decision of machine learning models while treating them to be a black-box. However, some approaches, such as those advocated in (Biran and McKeown 2014; Hendricks et al. 2016), are able to support their decisions with explanations justifying them with evidence from visual and semantic cues.

To extend the concept of explainability to video activity interpretation, we consider an explanation to be a description that explains and justifies the rationale of a model's decision process. In addition to providing justification with respect to both feature-level evidence, we also focus on explaining how the semantic correlations are established among concepts that make up an activity (actions and objects). In open, uncontrolled environments, establishing justifiable semantic correlation is integral to a model's success since the training data may not always be representative of all viable activities that one may encounter. It should be noted that we consider an explanation to be both introspective as well as retrospective.

A model's ability to provide sufficient justification for its decision requires in-depth knowledge about various concepts and the relationships that they share with other concepts. This use of prior knowledge can be considered to be analogous to how humans correlate the presence of certain concepts to aid in the current task. For example, in medical diagnosis (Ledley, Lusted, and Ledley 1959), it has been noted that the reasoning process used by doctors requires the establishment of correlation between symptoms (logical concepts) and probabilities to aid their diagnosis. Each symptom adds a certain value to the overall diagnosis and hence influences the direction of the reasoning process. This prior knowledge can be particularly helpful in identifying *how* two concepts can be related and *why* that relationship can contribute to the overall goal of the model.

In this paper, we propose a novel framework that leverages Grenander's Pattern Theory structures (Grenander

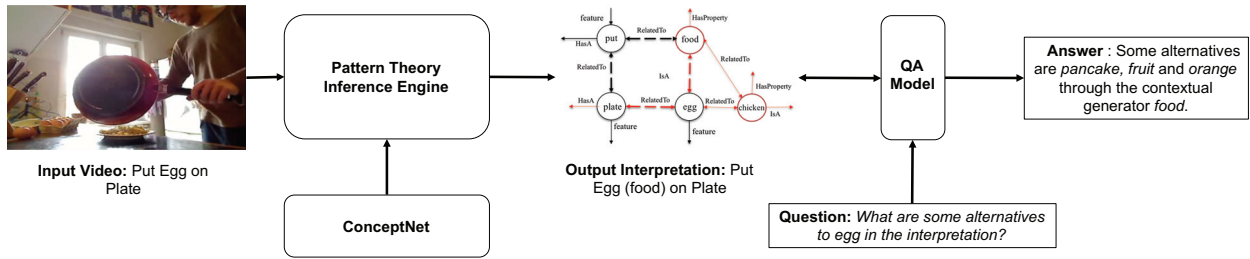


Figure 1: Overall architecture Deep learning or machine learning-based approaches hypothesize multiple object and action labels. Pattern theory formalism disambiguates knowledge using ConceptNet to generate an interpretation. An interactive agent then uses this as a source of knowledge for conversation about the inference process.

1996) to infer semantically coherent interpretations of video activity. An interpretation is defined as a semantically linked structure of concepts. It is an intermediate representation that can be considered to be the underlying source of knowledge for more expressive representations such as sentence-based descriptions and/or question and answers systems. In pattern theory language, concepts are represented by basic elements called generators with their semantic relationships represented by connections called bonds. Some concepts in this representation possess direct evidence from video, i.e. grounded concepts, while some are inferred concepts called contextualization cues. As defined by Gumperz (Gumperz 1992), primarily for linguistics, contextualization refers to the use of knowledge acquired from past experience to retrieve *presuppositions* required to maintain involvement in the current task. It has also been observed that providing contextualization cues often result in increase in acceptance of decisions made by automated systems (Herlocker, Konstan, and Riedl 2000; Martens and Provost 2013).

The overall architecture of the proposed approach is shown in Figure 1. Given an input video, individual, atomic concepts such as actions and objects are hypothesized using machine-learning or deep learning approaches. The resulting, multiple putative labels per object instance are then used to generate interpretations using an Markov Chain Monte-Carlo (MCMC) based simulated annealing process. The most likely interpretations are then used as the source of knowledge for generating explanations for human interaction via a dialog model.

The contribution of this paper is three-fold: (1) we are, to the best of our knowledge, among the first to address the issue of explainability in video activity interpretation; (2) the use of contextualization cues allow us to generate interpretations that is able to provide sufficient information to generate explanations at different levels of abstraction - from feature-level evidence (through grounded generators) to semantic relations (via bonds); and (3) we are able to show, through a dialog model, that the proposed framework is capable of generating explanations for its decision making process that is both introspective and retrospective.

Explainable Model for Video Interpretation

Grenander’s formalism allows us to express interpretations in an inherently explainable manner facilitating better hu-

man interaction. We begin with discussion about how concepts (such as actions and objects) are represented as *generators* and the different types of generators that can exist. We follow with discussion on how the detected concepts are grounded with semantic provenance using contextualization cues generated from a commonsense knowledge base known as ConceptNet (Liu and Singh 2004). We, then, follow with discussion on how generators are connected together using connections called *bonds* to form video interpretations called *configurations*. Finally, we end with discussion on the Monte-Carlo based inference process.

Symbolic Representation of Concepts

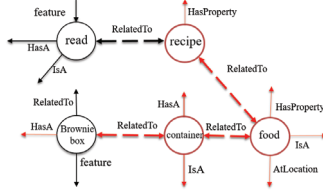
Following Grenander’s notations (Grenander 1996), we represent each concept using atomic components called generators $g_i \in G_S$ where G_S is called the generator space. The generator space represents a finite collection of all possible generators that can exist in a given environment.

The generator space (G_S) consists of three disjoint subsets that represent three kinds of generators - feature generators (F), grounded concept generators (G) and ungrounded context generators (C). Feature generators ($g_{f_1}, g_{f_2}, g_{f_3}, \dots, g_{f_q} \in F$) correspond to the features extracted from videos and are used to infer the presence of the basic concepts (actions and objects) called grounded concept generators ($\underline{g}_1, \underline{g}_2, \underline{g}_3, \dots, \underline{g}_k \in G$). Individual units of information that represent the background knowledge of these grounded concept generators are called ungrounded context generators ($\bar{g}_1, \bar{g}_2, \bar{g}_3, \dots, \bar{g}_q \in C$).

Each type of generator is a source of knowledge for generating explanations and hence contributes to the overall interpretation’s inherent explainability. For example, the feature generators allow the model to establish and express provenance for grounded concept generators in the actual input data. Hence, the model is able to provide direct video evidence for the presence of the grounded concept generators in the final configuration. The ungrounded context generators represent the additional, background knowledge that allow us to *semantically* correlate the presence of the grounded concept generators and hence help provide semantic justification for the presence of a concept in the final interpretation.



Input Video: Read brownie box



Top Interpretation: Read (container) (food) (recipe) brownie box

Q: How did you arrive at the interpretation?

A: There are two feature generators that are to be considered – CNN and CNNFlow. The grounded concept generators *read* and *brownie box* were derived from these feature generators with energies -0.03862 and -0.92633 respectively. Semantically, the concepts *read* and *brownie box* are connected through the contextual generators *container*, *food* and *recipe*.

Q: What are some alternatives to *read* in the interpretation?

A: Some alternatives are *take*, *pour* and *stir*.

Q: Why not the concept *brownie bag* instead of *brownie box* in the interpretation?

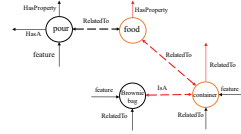
A: The concept generator *brownie bag* can be explained through the feature generator *CNN*. No semantic relationships can be established between concept generators *brownie bag* and *read*.

Q: What are some alternatives to the interpretation?

A: Some alternatives are:



Top Interpretation: Take (container) (food) brownie box



Top Interpretation: Pour (container) (food) brownie box



Top Interpretation: Read (container) (food) brownie bag



Top Interpretation: Take (container) (food) bowl

Figure 2: An illustration of an example interaction with the proposed model when provided with a video with groundtruth "Read brownie box". The model is able to provide a walk-through of its inference process and justifies the presence of each concept in its final interpretation at both data and semantic levels. Note: This is a visualization of the answers generated by the model.

Constructing Contextualization Cues

In the context of video activity recognition, we propose the use of a commonsense knowledge base as a source of contextualization cues for establishing semantic relationships among concepts. ConceptNet, proposed by Liu and Singh (Liu and Singh 2004) and expanded to ConceptNet5 (Speer and Havasi 2013), is a knowledge source that maps concepts and their semantic relationships in a traversable semantic network structure. Spanning more than 3 million concepts, the ConceptNet framework serves as a source of cross-domain semantic information from general human knowledge while supporting commonsense knowledge as expressed by humans in natural language. Technically, it encodes and expresses knowledge in a hypergraph, with the nodes representing concepts and edges representing semantic assertions.

There are more than 25 relations (also referred to as assertions) by which the different nodes are connected, with each of these relations contributing to the semantic relationship between the two concepts such as *HasProperty*, *IsA*, and *RelatedTo* to name a few. The validity of each assertion in ConceptNet is quantified by a weighted score and is representative of the semantic relation between concepts. Positive values indicates assertions and negative values indicates the opposite.

Expressing Semantic Relationships

Each generator g_i has a fixed number of bonds called the *arity* of a generator ($w(g_i) \forall g_i \in G_S$). These bonds symbolic representations of the semantic relationships shared between generators. Bonds are differentiated at a structural level by the direction of information flow that they represent - *in-bonds* and *out-bonds*. Each bond is identified by a unique coordinate and bond value such that the j^{th} bond of

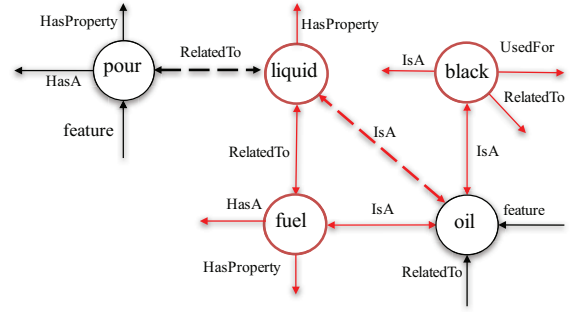


Figure 3: Representation of an interpretation using pattern theory. Black circles are generators that represent grounded concepts and red generators represent ungrounded concepts i.e. contextualization cues. The red links represent contextual bonds. Dashed links represent the optimal relationship between concepts.

a generator $g_i \in G_S$ is denoted as $\beta_{dir}^j(g_i)$, where dir denotes the direction of the bond. A bond is said to be *open* if it is not connected to another generator through a complementary bond. For example, in Figure 3 there exist a bonded generator pair $\{pour \text{ and } liquid\}$. The bonds representing *HasProperty* and *HasA* are *open*, whereas the bond labeled *RelatedTo* represents a *closed* bond between the generators "pour" and "liquid".

Types There exist two types of bonds - *semantic* bonds and *support* bonds. Each *closed* bond (both semantic and support) is a symbolic representation of a concept's ties to the interpretation and are used as guiding cues for generating explanations. The direction of *semantic* bonds signify the semantics of a concept and the type of relation-

ship a particular generator shares with its bonded generator or concept. These bonds are analogous to the assertions present in the ConceptNet framework. For example, Figure 3 illustrates an example configuration with *pour*, *oil*, *liquid*, etc. representing generators and the connections between them, given by *RelatedTo*, *IsA*, etc., representing the semantic bonds. *Support bonds* connect (grounded) concept generators to feature generators and are representative of direct image evidence for the grounded concept generator. These bonds are quantified using confidence scores from classification models.

Quantification The bonds between the generators are quantified using the strength of the semantic relationships between generators. This allows to quantify the amount of contribution the generator provides to the interpretation. The bond energy is quantified by the bond energy function:

$$a(\beta'(g_i), \beta''(g_j)) = q(g_i, g_j) \tanh(f(g_i, g_j)). \quad (1)$$

where $f(\cdot)$ is the weight associated with the relation in ConceptNet between concepts g_i and g_j through their respective bonds β' and β'' . The tanh function normalizes the score output by $f(\cdot)$ to range from -1 to 1. $q(g_i, g_j)$ weights the score output by the \tanh function according to the bond connection type (e.g., semantic or support) β' and β'' formed.

Constructing Interpretations

Generators can be combined together through their local bond structures to form composite structures called *configurations* c , which, in our case, represent semantic interpretations of video activities. Each configuration has an underlying graph topology, specified by a connector graph σ . The set of all feasible connector graphs σ is denoted by Σ , also known as the connection type. Formally, a configuration c is a connector graph σ whose sites $1, 2, \dots, n$ are populated by a collection of generators g_1, g_2, \dots, g_n expressed as $\sigma(g_1, g_2, \dots, g_i)$. The collection of generators g_1, g_2, \dots, g_i represents the semantic content of a given configuration c . For example, the collection of generators from the configuration in Figure 3 gives rise to the semantic content “*pour oil (liquid) (fuel) (black)*”.

Probability The probability of a particular configuration c is determined by its energy as given by the relation

$$P(c) \propto e^{-E(c)} \quad (2)$$

where $E(c)$ represents the total energy of the configuration c . The energy $E(c)$ of a configuration c is the sum of the bond energies formed by the bond connections that combine the generators in the configuration, as described in Equation 1.

$$E(c) = - \sum_{(\beta', \beta'') \in c} a(\beta'(g_i), \beta''(g_j)) + k \sum_{\bar{g}_i \in G'} \sum_{\beta_{out}^j \in \bar{g}_i} [D(\beta_{out}^j(\bar{g}_i))] \quad (3)$$

where G' is a collection of ungrounded contextual generators present in the configuration c , β_{out} represents each *out-bond* of each generator g_i and $D(\cdot)$ returns is function that

true of the given bond is open. k is an arbitrary constant that quantifies the extent of the detrimental effect that the ungrounded context generators have on the quality of the interpretation.

Inference

Searching for the best semantic description of a video involves minimizing the energy function $E(c)$ and represents the inference process. The solution space spanned by the generator space is very large as both the number of generators and structures can be variable. For example, the combination of a single connector graph σ and a generator space G_S give rise to a space of feasible configurations $C(\sigma)$. While the structure of the configurations $c \in C(\sigma)$ is identical, their semantic content is varied due to the different assignments of generators to the sites of a connector graph σ . A feasible optimization solution for such exponentially large space, is to use a sampling strategy. We follow the work in (de Souza et al. 2016) and employ a Markov Chain Monte Carlo (MCMC) based simulated annealing process. The MCMC based simulation method requires two types of proposal functions - global and local proposal functions.

A connector graph σ is given by a global proposal function which makes structural changes to the configuration that are reflected as jumps from a subspace to another. A swapping transformation is applied to switch the generators within a configuration to change of semantic content of a given configuration c . This results in a new configuration c' , thus constituting a move in the configuration space $C(\sigma)$.

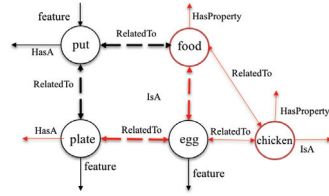
Generating Explanations

We are able to walk through the decision making process and express *why* it arrived at the interpretation as the most likely one. This allows the human to understand the reasoning behind the interpretation and provide a deeper understanding about how the interpretation is a viable explanation for the given video activity. In our current implementation, we allow for six questions that can be used to gain explanatory insight. A more general framework will be focus of future work. These questions allow us to evaluate the model’s ability to *justify* its decision as well as enhance its ability to interact with humans. The six questions are designed to (1) build an understanding of how the model is able to infer interpretations for a given video and enable us to walk through each aspect of its interpretation, (2) understand its drawbacks and possibly address them, and (3) use this interaction to learn new semantic relations if they do not already exist or strengthen the existing relationships for better performance in subsequent interpretations. They are the following.

- *How did you arrive at the interpretation?:* The model walks through the inference process starting at the feature level. The response enables the human to understand how each factor contributes to the interpretation and determine the point of failure (if any) for improving the performance.
- *What are alternatives to the interpretation?:* The answer to this question provides alternative interpretations. This



Input Video: Put egg on plate



Top Interpretation: Put (chicken) egg (food) on plate

Q: How did you arrive at the interpretation?

A: There are three features that are to be considered – HOF, HOG1 and HOG2. The grounded concept generators *put*, *egg*, and *plate* were derived from these feature generators with energies -0.76159, -0.87547 and -0.83807 respectively. Semantically, the concepts *put* and *plate* were connected through the *RelatedTo* bond. The concepts *put* and *egg* were connected through the contextual generator *food*. The concepts *plate* and *egg* were connected through the contextual generator *egg*.

Q: What are some alternatives to *egg* in the interpretation?

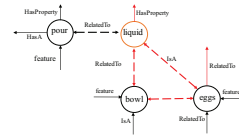
A: Some alternatives are *pancake*, *fruit* and *orange* through the contextual generator *food*.

Q: Why is the concept *egg* in the interpretation?

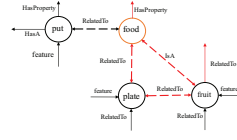
A: The concept *egg* is a possible explanation for the feature *HOG2* with energy -0.87547. Semantically, it is related to the other grounded generators. It is directly related to *plate*. It is connected to *put* through the contextual generator *food*.

Q: What are some alternatives to the interpretation?

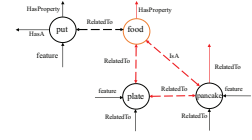
A: Some alternatives are:



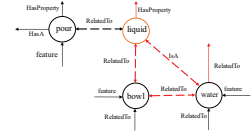
Alternate Interpretation: Pour eggs(liquid) to bowl



Alternate Interpretation: Put fruit (food) on plate



Alternate Interpretation: Put pancake (food) on plate



Alternate Interpretation: Pour water (liquid) to bowl

Figure 4: An illustration of an example interaction with the proposed model is shown when provided with a video "Put egg on plate". When prompted about the presence of a concept in its interpretation, the model is able to provide justification for presence of the concept at both data-based and semantic levels. It is also able to provide alternatives for both concepts and whole interpretations. Note: This is a visualization of the answers generated by the model.

allows the human to pick the best possible interpretation from the model. In a critical scenarios, a human may need to choose an alternative interpretation rather than blindly trusting the model's top prediction.

- **Why <concept> in the interpretation?:** The model looks for cues to justify the presence of a concept within its final interpretation. This provides a detailed justification for including a concept in the interpretations at both levels of abstraction - feature level and semantic level.
- **What alternatives to <concept> in the interpretation?:** To answer this, we walk through the inference process to bring alternatives to the specific concept in the interpretation. This allows for better understanding of the inference process while providing an ideal point of interaction for understanding the model's capability to semantically associate different concepts in a coherent manner.
- **Why not <concept1> instead of <concept1>?:** To answer this, we have to reason about alternatives. This interaction allows us to understand how the semantics influence its inference process.
- **The correct interpretation is <interpretation>. Why did you not get there?:** This prompts the model to continue reasoning about its inference process and provide a concise argument about its choices.

Considered together, these questions cover various aspects about the decision making process and explain the rationale behind the output. An important observation to be noted is that these questions require the model to be able to relate concepts together beyond what may be visible in the video data and/or training data. Hence models is able to generate semantically coherent interpretations as well as provide semantic justification for the presence of concepts beyond feature-level evidence.

Understanding Provenance of Concepts

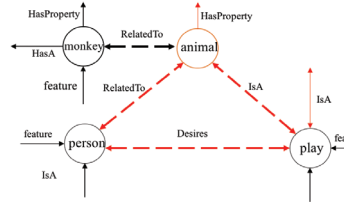
When dealing with complex video activities, understanding the rationale behind the presence of individual concepts in an interpretation is essential and requires meaningful explanations that explain the provenance of concepts. Often, it involves explanations that provide justification for the concepts that are both grounded with direct evidence from data as well as meaningful explanations that are not obvious in the video alone. In the proposed framework, interpretations are used as a source of knowledge to generate explanations for a concept's provenance. Direct data evidence for the presence of the concept is provided through the presence of feature generators while semantic justification is derived using the bonds connectign ungrounded context generators. For example, consider the interpretation in Figure 3 whose semantic content is *Pour oil*. The presence of the concept *oil* can be explained through the presence of its corresponding feature generator connected through the bond labeled *feature* as well as the ungrounded context generator *liquid*. Hence the resulting semantic explanation can be constructed as "*Oil can be poured because it is a liquid.*"

Understanding the Overall Interpretation

Another important aspect of explainability in such models is the ability of the model to express its decision making process that provides sufficient details to understand its rationale. This is essential, especially under critical circumstances where even a minor error could have disastrous consequences such as in medical diagnosis and surveillance. Such explanations, through meaningful interactions, can aid in understanding the systems overall strengths and weaknesses and convey an understanding of how the system will behave in the future. This could, perhaps, even be extended by the user to anticipate and possibly correct the systems



Input Video: Monkey is karate kicking at a man's gloved hand



Top Interpretation: A monkey is playing with a person

Q: How did you arrive at the interpretation?

A: There are three feature generators that are to be considered – CNN1, CNN2 and CNNFlow. The grounded concept generators *person*, *play* and *monkey* were derived from these feature generators with energies - 0.07720, -0.01929 and -0.511584 respectively. Semantically, the concepts *play* and *monkey* are connected through the contextual generator *animal*. *Person* and *play* are connected through the bond *Desires*.

Q: Correct interpretation is *Monkey fighting with person*. Why did you not get there?

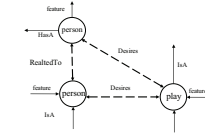
No semantic relationships could be established between concept generators *monkey*, *person* and *fight*. Hence, it has higher energy compared to “*Monkey is playing with person*” and resulting in lower probability of selection.

Q: Why not the concept *fight* instead of *play* in the interpretation?

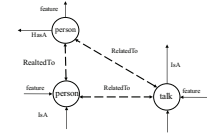
A: The concept generator *fight* can be explained through the feature generator *CNNFlow*. No semantic relationships could be established between concept generators *monkey*, *person* and *fight*.

Q: What are some alternatives to the interpretation?

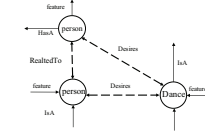
A: Some alternatives are:



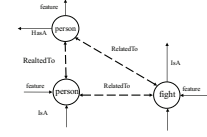
Alternate Interpretation: A person is playing with another person



Alternate Interpretation: A person is talking with another person



Alternate Interpretation: A person is dancing with another person



Alternate Interpretation: A person is fighting with another person

Figure 5: An illustration of an interaction with the proposed model when provided with a video “A monkey is fighting with a person”. It is to be noted that, when prompted, the model was able to provide an explanation that best describes the rationale behind its decision. It was also able to provide alternatives for a specific concept as well as the whole interpretation. Note: This is a visualization of the answers generated by the model.

mistakes. For example, consider Figure 2, where the model is prompted to explain its decision making process in the first interaction. It can be seen that the model begins with the factors that were considered in generating the interpretation - namely the feature generators CNN and CNNFlow and continues with the labels chosen to represent these feature generators in the final interpretation as grounded concept generators as well as expressing the confidence levels in its choice of labels. The model then is also able to justify the interpretation’s overall meaning through the presence of the ungrounded context generators *container*, *food* and *recipe*; thus covering all aspect’s of the model’s inference process.

Handling What-Ifs

Perhaps the most important aspect of explainability is a model’s ability to handle “What-if” scenarios posed by the user. As the final decision maker, the human may have some insight that the model does not possess such as intuition and experience. The model must be able to handle such queries and justify its inference process based on its experience and the resulting knowledge. For example, while an interpretation made by the model may hold semantic meaning, the context may not be correct and hence a exchange of concepts is required for better performance. This requires the model to have a deep understanding of the domain concepts and their applicability in the current interpretation. One such explanation is shown in Figure 5 where the prompt by the user posed an alternative concept *fight* in the place of the existing concept *play*. The model was able to reason through the semantic relationships in ConceptNet and able to justify its choice due to the lack of semantic concepts that allowed for semantic relationships with the other grounded concept generators *monkey* and *person*. Another example is shown in

Figure 2 where the model is able to reason about its failure to establish semantic relationships due to the lack of prior knowledge. It is important to note that such interactions can easily be extended into a form of active learning model that successfully transfers knowledge from the human user to its existing knowledge base.

Conclusion and Future Work

In this paper, we explore the aspect of explainability in intelligent agents that generate interpretations of multimedia data through the inherent nature of pattern theory structures and contextualization cues constructed from ConceptNet. We have so far evaluated the outputs on the Breakfast Actions dataset for over 5000 videos, but mostly qualitatively and visually, ourselves. We plan to conduct a structured study of the quality of the Q&A using human subjects. We demonstrate that the proposed approach naturally captures the semantics in ConceptNet to infer rich interpretations.

Acknowledgement

This research was supported in part by NSF grants IIS 1217676 and CNS-1513126. The authors would also like to thank Daniel Sawyer for his invaluable insights during discussion.

References

- Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and MÄžller, K.-R. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11(Jun):1803–1831.
- Biran, O., and McKeown, K. 2014. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML*, volume 2014.

- Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. ACM.
- Core, M. G.; Lane, H. C.; Van Lent, M.; Gomboc, D.; Solomon, S.; and Rosenberg, M. 2006. Building explainable artificial intelligence systems. In *AAAI*, 1766–1773.
- de Souza, F. D.; Sarkar, S.; Srivastava, A.; and Su, J. 2016. Spatially coherent interpretations of videos using pattern theory. *International Journal on Computer Vision* 1–21.
- Grenander, U. 1996. *Elements of pattern theory*. JHU Press.
- Gumperz, J. J. 1992. Contextualization and understanding. *Rethinking context: Language as an interactive phenomenon* 11:229–252.
- Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *European Conference on Computer Vision*, 3–19. Springer.
- Herlocker, J. L.; Konstan, J. A.; and Riedl, J. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 241–250. ACM.
- Junior, J. C. S. J.; Musse, S. R.; and Jung, C. R. 2010. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine* 27(5):66–77.
- Kheradpisheh, S. R.; Ghodrati, M.; Ganjtabesh, M.; and Masquelier, T. 2016. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports* 6:32672.
- Lane, H. C.; Core, M. G.; Van Lent, M.; Solomon, S.; and Gomboc, D. 2005. Explainable artificial intelligence for training and tutoring. Technical report, DTIC Document.
- Ledley, S.; Lusted, L. B.; and Ledley, R. S. 1959. Reasoning foundations of medical diagnosis. In *Science*. Citeseer.
- Linder, N.; Turkki, R.; Walliander, M.; Mårtensson, A.; Diwan, V.; Rahtu, E.; Pietikäinen, M.; Lundin, M.; and Lundin, J. 2014. A malaria diagnostic tool based on computer vision screening and visualization of plasmodium falciparum candidate areas in digitized blood smears. *PLoS One* 9(8):e104855.
- Liu, H., and Singh, P. 2004. Conceptnet: a practical common-sense reasoning tool-kit. *BT Technology Journal* 22(4):211–226.
- Lomas, M.; Chevalier, R.; Cross II, E. V.; Garrett, R. C.; Hoare, J.; and Kopack, M. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 187–188. ACM.
- Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 1975–1981. IEEE.
- Martens, D., and Provost, F. 2013. Explaining data-driven document classifications.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.
- Shortliffe, E. H., and Buchanan, B. G. 1975. A model of inexact reasoning in medicine. *Mathematical biosciences* 23(3-4):351–379.
- Speer, R., and Havasi, C. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*. Springer. 161–176.