Reverse Engineering Creativity into Interpretable Neural Networks

Marilena Oita

The Swiss AI Lab IDSIA, SUPSI, USI
Lugano, Swizerland
marilena@idsia.ch

Abstract. In the field of AI the ultimate goal is to achieve generic intelligence, also called "true AI", but which depends on the successful enablement of imagination and creativity in artificial agents. To address this problem, this paper presents a novel deep learning framework for creativity, called INNGenuity.
Pursuing an interdisciplinary implementation of creativity conditions, INNGenuity aims at the resolution of the various flaws of current AI learning architectures, which stem from the opacity of their models. Inspired by the neuroanatomy of the brain during creative cognition, the proposed framework's hybrid architecture blends both symbolic and connectionist AI, inline with Minsky's "society of mind". At its core, semantic gates are designed to facilitate an input/output flow of semantic structures and enable the usage of aligning mechanisms between neural activation clusters and semantic graphs. Having as goal alignment maximization, such a system would enable interpretability through the creation of labeled patterns of computation, and propose unaligned but relevant computation patterns as novel and useful, therefore creative.

Keywords: creativity, neural networks, imagination, semantic networks, knowledge, interpretability, neural architecture

1 Introduction

Creative cognition, or *useful imagination*, represents the missing functionality which complements the machine *learning*, towards achieving a sense of flow, and uniqueness necessary to generic AI [38]. The ability to generate novel and useful ideas, i.e., solutions, is a key driver of culture creation and human evolution, and of inestimable value in today's world.

The broadly adopted definition of creativity refers to the process by which cognitive systems instigate the genesis of novelty, which is goal-appropriate [4]. Humans being the only agents known for their ability to achieve creative outcomes with respect to the above definition, it goes per se that, if we want to instigate the act of creation in machines, we need to understand and mimic the way humans operate. Reverse engineering the process of creativity and putting together latest discoveries and insights from neuroscience, cognitive science, and artificial intelligence, is fundamental in architecting the *creative gene*.

Creativity is often described in terms of novelty, but only through the process of decoding (i.e., interpreting) an outcome can we qualify it as novel or useful. The *new* can

therefore be assessed only with respect to a reference, which is our current understanding, i.e., knowledge of the world. Following the intuition that a new solution cannot be found in the same "type of thinking" [1], the process of creation needs a *continuous exchange* with a greater context than that of the environment in which the problem exist. Nevertheless, an AI architecture that takes into account this necessity has not been yet considered, which results in the lack of means towards *generic AI*.

Various branches of philosophy, psychology and evolutionary biology agree that creative ideas should not be just new and unexpected, but are also successful at providing solutions that are both valuable and *efficient*. Creativity is a computationally efficient process [28]. Driven by optimization, agents use creative cognition to build incrementally on 'simpler' (after understanding observations of reality, fewer computational resources are needed for an agent to encode/decode) patterns. When shared and adopted globally, creative solutions result in knowledge. Creativity feeds on what we already know, constantly learn, and occurs in the information integration process. All knowledge is the result of past creativity, and a binding material for future one.

As opposed to learning, which is conceptual acquisition, the core characteristic of creativity is conceptual expansion: the ability to widen one's conceptual structures to include unusual or *novel associations* [32]. This expansion implies that a new relation is created between the "old" and the "new", and in order to perform this integration step we need to understand the patterns of the creative output. Following this reasoning, a creative process needs to be interpretable.

Besides interpretability, for which a possible solution is presented in section 4, many interdisciplinary advancements are ready to be used to create a 'society of mind' [22]: the neuroanatomy of creativity which has been recently revealed using RMI technology in neuroscience [27], goal-oriented measures which assess the success of generated ideas [34], neural networks which pay attention to semantic cues, successfully reuse computation and share a conceptual space [39], semantic graphs which become larger every day through the use of Linked Data and machine learning at scale.

Semantic technologies and neural networks have been rivals at approaching 'true AI', with only few, but memorable influences like Marvin Minsky, aiming at their collaboration [30]. Neuroscience also argues that creative thinking emerges through the dynamic interplay between various, functionally diverse, components, Providing an architecture which blends symbolic and connectionist approaches in a suite of cooperating methods is the goal of the INNGenuity framework.

The structure of this paper continues as follows: Section 2 proposes pluridisciplinary perspectives regarding the creation and communication of knowledge, and its link to creativity. Section 3 presents the complementary aspects of symbolic and connectionist approaches to AI when considering a creative purpose. The INNGenuity framework's components and envisioned implementation are described in Section 4, while Section 5 outlines conclusions and future work.

[1] https://www.brainyquote.com/quotes/albert_einstein_385842

2 Reverse-engineering Creativity

Forging artificial creativity is dependent on first, our ability as a society to renounce at the belief that humans are particularly special because of their unique ability to exerce creative cognition. Second, we need to overcome the supposed complexity which makes the field of creativity largely unexplored by scientists in comparison with its value.

Following the ability to transfer information from memory to external digital resources, to externalize and perform computation at scale using the cloud, enabling imagination in artificial agents is the ultimate resource to be developed in order to further expand our ability to understand the world.

At a philosophical level, the universe itself is a mass creation, whose components all share, at different intensities, and varying qualities, the property of being creative. Most often built as survival strategies, and explained by evolutionary biology as such [19], the creativity mechanisms ensure not only the persistence of information through precise transmission (e.g., genes), but also its expansion either in sheer quantity, or quality (i.e., sophistication) [20].

Humans in particular have evolved to decompose observations of reality into abstractions, or structured mental representations [1], which can be used as building blocks [2] in the construction of more sophisticated abstractions [3], which aim at re-constructing the "reality". The world, as we perceive it, is partially creativity understood (or, "decoded"), and partially creativity "encoded"(not yet revealed). Therefore, our goal, as "decoding" machines [33], is to understand it. Indeed, humans seem to be drawn, as soon as the survival conditions have been (even loosely) satisfied [10], towards the understanding of one-self and the world [29]. This understanding is made possible by the use of a communication framework i.e., a collection of conventional codes to be shared between agents, which ensures the endurance and expansion of the global creative outcome. Indeed, recent advances in neuroscience [8; 9; 20; 26] agree on the fact that language first evolved as a cognitive tool, and only afterward was externalized for information transfer.

Communicating the creative outcome, whether it is mundane or extraordinary, is essential to its assessment and adoption. For practical reasons, in order for the outcome to be useful to the users of the creation, artificial agents need to perform the resolution of the unknown *in a human interpretable way*.

In this paper, the computation framework typically used in cutting-edge AI, deep neural networks, is complemented by the communication framework represented by semantic networks. In the process of rendering black-box models interpretable using semantics, the neural network's shortcomings in relation to creativity: opacity, high computational needs, and narrow focus are also reduced.

3 Interpretable Neural Networks

Being able to learn nonlinear latent representations through the activation functions, while being highly effective in capturing local relationships, neural networks have been successfully leveraging powerful computational resources and big data. Choosing not to 'believe without questioning', and recalling that model 'silence' can create monsters,

many scientists denounce the black-box model of the world which is presented to us, and endeavor to incorporate side information (i.e., semantics), but face architectural limits.

Rendering machine learning models interpretable would have huge societal impacts: it would not only enable a wider adoption of AI by domain-critical systems, such as medical, but also for security and ethical reasons.

The symbol-oriented community of AI supports models which are self-describing, but alone too rigid and specialized. Semantic networks are graphs which describe a domain using entities, concepts and relations, and have built-in expansion mechanisms which ensure that the "possibly new" is integrated to the "existing" in a consistent manner. Modelling a series of facts, in the form of triples (Subject Predicate Object), semantic graphs such as ontologies represent a flexibile way of reasoning on a domain, either generic (law, chemistry) or specific (e.g., state law regulation, molecule interactions etc.).

Common sense knowledge bases like CyC focus on things that rarely get written down or said, providing a causal understanding of the world we live in, which would come at hand to artificial agents. In addition, generic ontologies built from Wikipedia like Freebase, DBPedia, Yago etc., or built by experts like Wordnet, provide a vast cultural coverage. This knowledge has been successfully leveraged by search and QA systems, e.g., Watson AI system, or the Google Graph.

Semantic technologies have nevertheless the shortcoming of lacking a computation framework that sustains the acquisition of new knowledge, and efficiently updates the existent one. Building the facts of an ontology is usually a semi-automatic or manual, costly process involving expert validation. The expert needs to use her own computation node (ie., the brain) to compensate the lack of an automatic, unsupervised framework.

Since a semantic network is built for reuse, the facts need to be true, i.e., logically provable. In reaction to this excessive care for consistency, the connectionist approaches try to avoid the high-maintenance costs and build architectures endowed with as little knowledge as possible. Nevertheless, the results are integrating data biases [2] which are not obvious and dangerous in the lack of interpretability. By stripping away knowledge from the computation, relying on pure numerical signals, and building black-box models of the world, connectionist approaches also disregard ways in which models could be improved and reused.

The fundamental flaws which create an algorithmic consumerism in our society, and represent a risk factor in many real-world applications are: 1. model opacity, 2. high resource consumption, and 3. narrow focus. The hypothesis that using experience in the form of data, when large, it allows for reaching broader conclusions reaches its limits: a ”plateau” of efficacy mitigated only by ever-growing resources. This status-quo favors the passive usage of computing resources by those who can afford them, instead of a compute-encode-reuse strategy which is smarter in terms of optimization. Recently, the fields of transfer and multi-task learning try to mitigate this situation, but their concern is more related to reusing encoded data patterns, but not expliciting them, therefore interpretability continues to be an issue.

In addition, since the generalization provided by an NN solves problems of ”the same type”, by definition, this narrow perimeter cannot allow the necessary conceptual

[2] https://web.media.mit.edu/ minsky/papers/SymbolicVs.Connectionist.html

leap in forging creative solutions. Although narrow focus could be considered a quality, it maintains the learning in a space in which creation could never be instigated. Currently, once an agent trained with NN has learned to play (e.g., Go), it will beat anyone at that task, but cannot do anything else. Additionally, the model lacks all means to understand what is doing, and in relation to what. This is against the generic AI dream, which states that an agent should be able to perform multiple tasks without being reprogrammed.

Attention added to NN has been successful in a number of applications, as a step towards coping with the noisy data problem by identifying relevant parts of the input for the prediction task. Recently, the Transformer architecture [36] claims better accuracies than LSTMs. Attention can only optimize the learning, and although it does not provide the necessary conditions for creativity, it is a proof that we can alter the architecture of neural networks in various beneficial ways. If attention is used as a mechanism to dynamically inject relevant external knowledge into the computation, then it becomes a bridge for the ideas presented in this paper.

4 The INNGenuity Architecture

The goal of this work is to introduce a variant NN architecture with in/out access to semantics through specialized gates. Its structure contains three modules, inspired by recent advances in neuroscience concerning the anatomy of creative cognition.

Over the past years, an important effort in neuroscience has been pursued to localize the creativity in the brain. INNGenuity framework aims to mimick these reverse-engineered processes concerning the brain circuitries (or hubs) underlying creative thought.

4.1 The anatomy of creative cognition

The human brain is recognized to function in a manner consistent with the notion of ”hubs” [18], communicating regions of the brain which have built-in mechanisms optimizing the information transfer [23], even across long distances [6]. These are:

The Imagination Hub is involved in ’constructing dynamic mental simulations based on past experiences, thinking about the future, and generally imagining alternative perspectives and scenarios to the present’ [17]. It is represented in the INNGenuity framework by typical connectionist approaches, namely NNs, but whose functioning has been semantically biased in an automated way.

The Salience Hub constantly monitors both external events and the internal stream of data, while giving priority to whatever information is most salient to solving the task at hand [3]. It is represented in the INNGenuity framework by the IN/OUT control mechanism possible though semantic gates, introduced further. This models the perception and attention modules of the brain in the form of a sensitivity to, or prioritary treatment of relevant observations.

[3] https://med.stanford.edu/content/dam/sm/scsnl/documents/Menon_Salience_Network_15.pdf

The Executive Hub is active when engaging in reasoning that puts heavy demands on the memory. Represented in the INNGenuity framework by semantic technologies, it solves the following: (i) encoding: finding similarities between the NN patterns' *structure* and the topical semantic graph which is proper to the domain of computation; (ii) decoding: maps meaning (i.e., labels) to encodings using an interpretation scheme, made of a vocabulary of symbols and the relations between them; (iii) abductive reasoning: combines the explanations obtained through the output semantic gates of the NN, and assesses which parts of the result are novel by comparing them with the *global* knowledge (i.e., alignment with a generic semantic graph), and the goal, respectively.

4.2 Semantic gates motivation and purpose

Knowing that creation comes with the pursue of a 'different resolution path than the expected one' [24], INNGenuity introduces semantic biases in order to create meaningful divergence from initial representations.

Adding special gates to NN is not new. The forget gate has introduced to LSTMs, but generally, the gates purpose is to optionally let information through, since they control access to memory cells. In theory, semantic gates can be added to any variant of NNs, for instance Long Short-Term Memory (LSTM) networks.

Current variants of LSTMs have the ability to read, write and erase data from the memory cells, but these data are static to the learning process. An alternative idea promoted in this work is to dynamically bias the input data at every step of the computation with relevant knowledge proper to that current layer level of abstraction. Besides its positive contribution to expanding the solution search space, access to background knowledge opens the door to a larger context that the local one, and has been endorsed as fundamental for creativity by both cognitive science [32], and neuroscience [12; 13]. LSTM networks learn to process data with complex and separated interdependencies, but restricted to the typical input-output settings, they are not capable of imagination, which requires a non-linear relationship to the data.

Starting from a typical NN, LSTM [11] for instance, in addition to the typical forget gate, *semantic gates*, denoted by k_{IN} and k_{OUT}, are introduced as in Figure 1.

The purpose of semantic gates to NN is twofold: 1. they operate towards a meaningful bias of the computation by adding contextual external knowledge to the internal data context of the NN's layers, and 2. they allow the output of the patterns identified at different layers of abstraction as clusters, i.e., graphs composed of impactful neural activation dependencies. This latter function is beneficial for interpretability: the clusters will pass through an alignment phase [21] with a topic semantic graph to explicit the computation.

The semantic gate implements the Salience Hub of attention and awareness. Allowing bidirectional communication between NN and semantic graphs, its logic is regulated by the aligning mechanisms of the Executive Hub, dealing with knowledge integration and meta-reasoning. Through the semantic gates, a bidirectional metadata flow provide the necessary conditions for creativity as introduced in [31]: 1. a mechanism for introducing variation (IN semantic bias), and 2. a mechanism for preserving and reproducing the selected variations (OUT persistent expression).

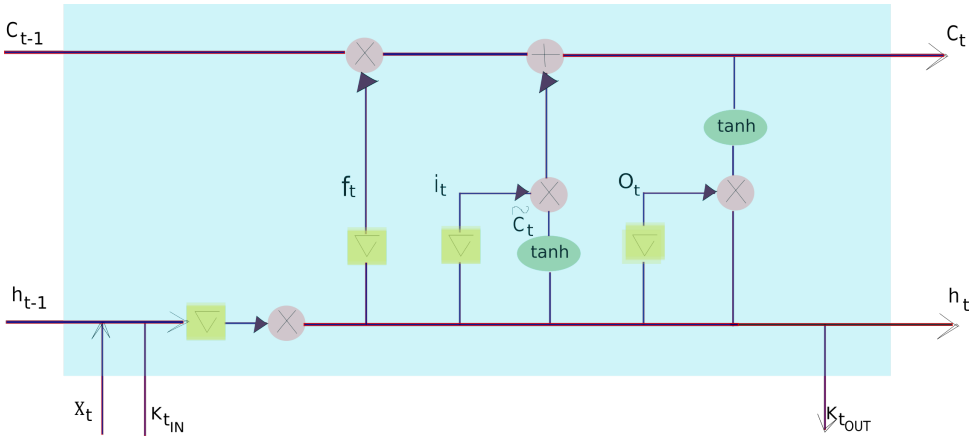

Fig. 1: Semantic gate IN/OUT contribution to memory cell

The exploratory and eliminatory aspects of the creativity process can be also mimicked [15]: 'a sequential back-and-forth begins with informed guesses and progresses to increasingly probable solutions', can be implemented in semantic networks through the mechanism of alignment towards the increasingly abstract. Semantics feature hierarchies are shown to be effective in boosting the accuracy, besides the interpretability of neural models [25].

A NN outlines a process in which the existence of labeled data (i.e., outcomes) is a condition for learning. In opposition, triggered when we face unknown situations, creativity implements an abductive reasoning process [14], i.e., an adaptive problem solving strategy. The abductive feature of the reasoning supposes that the outcome instances are not precisely defined, instead the outcome has to be 'invented', or *explicated using the most relevant facts*. In order to enable this abductive property of the creative process, but still using the NN's characteristic of an universal approximator, the 'focus' of the NN has to be relaxed. There are two ways of approaching that: 1. if outcomes exist, they will be conceptually expanded to a semantic graph; or 2. if outcomes are not precisely defined, then a semantic graph relevant to the problem to be solved (i.e., a topic) is projected using the access to a generic knowledge base.

4.3 Disentangling neural activations into clustered signals carrying semantics

Rendering an agent interpretable, is equivalent to transforming its black-box model into a self-explaining structure. This cannot be done a posteriori, therefore it needs to happen *while learning*.

Disentangling the underlying explanatory factors hidden in the observed environment is one important goal of Representation Learning [7]. Interpretability needs however an additional step involving attaching meta-information to the explanatory factors. In this paper, the approach to interpretability is employing semantic networks, which have the advantage of providing a self-describing graph structure.

Many successful learning systems benefit from prior knowledge about composition and structure, but the large majority are supervised, while this work describe an unsupervised approach.

The aim of hybrid neural systems is finding best ways to integrate both symbolic and connectionist approaches [16], but for that the connectionist mechanisms need to become more transparent. Only recently differentiable clustering algorithms have been applied on neural signals for the identification of objects and their interactions [35]. Intuitively, the relevant signals identified during computation correspond to object properties encoded. If the encoding (i.e., embedding) scheme is coherent and the transformations consistent, then the clustering would be able to outline the causal dependencies between the primitives of the compositional system, therefore semantically explicit the **isA** relation learned by the model.

For that, input and output data embeddings are used in the selection of more relevant information from a knowledge source, provided that embeddings of the knowledge graph have been computed in advance, e.g. see [37].

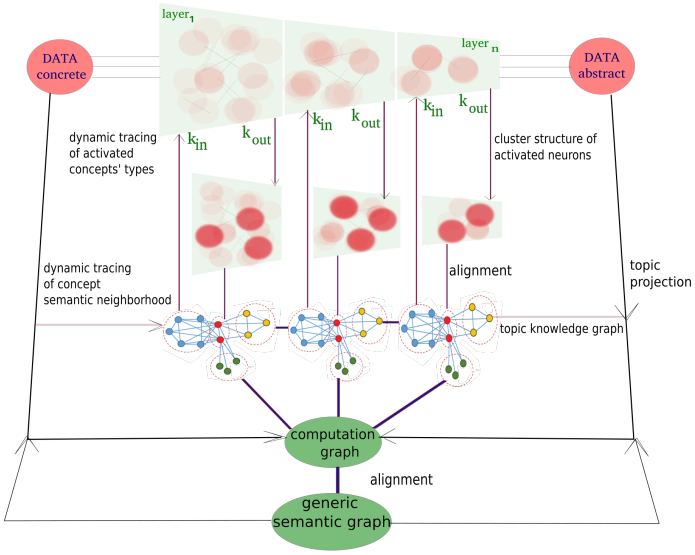

Fig. 2: The Clustering and Alignment Phases

As shown in Figure 2 the typical NN processing is complemented with a phase of clustering, and two phases of alignment with a knowledge graph. The optimization criteria for the NN is then to maximize alignment, in addition to the typical loss minimization. During NN computation the most impactful signals are being conceptually cached and clustered. The resulting clusters represent a graph (set of triples), which is outputted through the semantic gate of that layer. Then, the Executive Hub takes over, and in the

process of alignment with a topical semantic graph, the graph of clustered activations is being labelled with relevant concepts based on the structure similarities of the alignment candidates. Linking together labelled explanatory factors identified at every layer of abstraction is equivalent to rendering the computation steps explicit.

The use of semantic gates ensures that knowledge is used contextually, therefore the process diverges from a typical learning due to the introduction of a relevant variation. The results become imaginative because of their plausability, but only novel and useful ones will be considered creative. The assessment of these two properties is the goal of the next phase of alignment.

The *novel* and *useful* attributes of each of the triples (Subject Predicate Object) forming the labelled computation graph are measured. In order to do that, we need a second alignment and to make use of the impact weights of activated neurons. In this process, the generic knowledge base mimicks the 'culture' of that environment and serves as a reference for identifying new associations. Partially unaligned structures (triples) which are most impactful would be considered novel (and useful by the fact that they have been activated during computation), therefore by definition creative. Eventually, the culture, or collective memory would be constantly enriched by collaborative agents with new outcomes in the defined domain of operation, contribution which unifies at scale *views* on computations made by different agents.

In contrast with recent causal frameworks which aim at explaining the predictions of a NN model [5], INNGenuity dynamic flow allows a better flexibility and the possibility of designing a fully unsupervised process since both the clustering of patterns, and the semantic alignment are unsupervised.

From a symbolic perspective, INNGenuity designate a semantic network in which relations between concepts represent *computation nodes* (aka NN), and the relation label is a clarification of the computation node's purpose. From a connectionist perspective, INNGenuity designate a neural network in which the meaningful data patterns and their interactions are being transformed into a semantic graph which explicates the computation.

5 Discussion and Conclusions

Creative AI is the most human invention that we have the chance of pursuing. Using insights from neuroscience, the functions of a creativity framework are defined as the generation and assessment of novelty. This work argues that the generation of novelty needs a dynamic integration of meaningful conceptual structures (aka plausible biases) to memory cells. At the same time, the assessment of novelty needs interpretability. The goal of the INNGenuity architecture is to enable both interpretability and possible creative outputs. Its design instigates towards the concilliation of best AI practices for more impactful progress in this domain.

From a hybrid systems classification, INNGenuity outlines a connectionist symbol processing approach, a tightly coupled system in which knowledge and data are dynamically transferred and shared by the neural and symbolic components, via common internal structures: semantic gates.

The flow of meaningful conceptual structures at each NN layer of abstraction through intput semantic gates is thought as enabling means of 'inspiration' and context awareness. Purely guided by the recognition of patterns formed during computation, INNGenuity incorporates a differentiable clustering algorithm which outlines the structural dependencies between activated neurons. Activation clusters, which are assumed to carry hidden semantics and compositional logic, are further aligned with the rich and relevant knowledge of our world in the form of (nowadays pervasive) semantic graphs.

Knowledge, seen as the result of past computation and modelled by symbolic systems as semantic graphs, needs to be seemingly integrated into the live computation process for its huge potential in terms of resource optimization, but also for its unique capability to make unknown structural dependencies explicit by means of alignment.

Aligning pattern structures is a *decoding mechanism* which enables the creation of a labelled computation graph. This facilitates model understanding by humans, and communication with other artificial agents. The flow of such structures through output semantic gates is thought as enabling means of 'expression'. By means of semantic graphs alignment, patterns of computation get labeled, shared, and reused, resulting in knowledge being enriched incrementally and continuously in collaboration.

Eventually, unaligned parts of the semantic computation graph constitute the possible creative outcome. Since they represent structures activated during computation, these unaligned parts are considered useful, besides representing new associations.

One of the main advantages of the INNGenuity architecture is the integration of unsupervised approaches, such as alignment and clustering, thus allowing a high degree of automation.

A current condition towards the successful implementation of the INNGenuity approach is existence of efficient differentiable clustering algorithms operating on neural activations. Another limitation (but which will always exist due to our own limited ability to capture and express knowledge) is that achieving accurate and fine-grained interpretations depends on the quality and depth of the semantic graphs made available to an agent. In this direction, semantic gates are envisioned as a feedback loop between NN and semantic graphs, setting in which knowledge can grow and improve over time due to the controlled interaction between real-world data (sometimes also real-time), and the conceptual and compositional representations we have about these data.

As further work, the author aims at implementing a prototype of INNGenuity producing creative outcomes as a neural network operating in alignment with knowledge graphs. Its potential can be shown in applications in which creativity is more than encouraged: generative systems e.g., language models, dialogue systems and chatbots, or in other processes of assessing and boosting human creation, such as research.

Bibliography

[1] F. J. A. and P. Z. W. Minds without meaning: An essay on the content of concepts. *Cambridge, MA: The MIT Press*, 2015.

[2] K.-S. A. *Is creativity domain specific or domain general? Cases from normal and abnormal phenotypes*. Artificial Intelligence and Simulation of Behavior Quarterly 85, T.H. Dartnall, 1993.

[3] K.-S. A. *Digest of Beyond Modularity*. Behavioral and Brain Sciences, 17.4, 1994.

[4] R. M. A. and J. G. J. The standard definition of creativity. *Creativity Research Journal*, 2012.

[5] D. Alvarez-Melis and T. S. Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *CoRR*, abs/1707.01943, 2017.

[6] D. S. Bassett and B. E. Small-world brain networks. *Neuroscientist*, 12:512–523, 2006.

[7] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[8] R. A. C. Why language really is not a communication system: a cognitive view of language evolution. *Frontiers in Psychology*, pages 14–34, 2015.

[9] S. D. and W. D. Relevance: Communication and cognition. *Oxford: Basil Blackwel*, 1995.

[10] R. Dawkins. *The selfish gene*. 1941-(1989).

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[12] J. Hummel and K. Holyoak. A symbolic-connectionist theory of relational inference and generalization. 110:220–64, 05 2003.

[13] J. E. Hummel and K. J. Holyoak. Distributed representations of structure: A theory of analogical access and mapping. 104:427–466, 1997.

[14] R. Jung, B. Mead, J. Carrasco, and R. Flores. The structure of creative cognition in the human brain. *Frontiers in Human Neuroscience*, 7:330, 2013.

[15] S. D. K. Creative problem solving as sequential bvsr: exploration (total ignorance) versus elimination (informed guess). *Thinking Skills Creativity 8*, 2013.

[16] S. W. Kenneth McGarry and J. MacIntyre. Hybrid neural systems: From simple coupling to fully integrated neural networks. *Neural Computing Surveys*, pages 62–93, 1999.

[17] B. R. L., A.-H. J. R., and S. D. L. The brain's default network: anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci. 1124*, 2008.

[18] B. S. L. and M. V. Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.*, 2010.

[19] S. B. K. Liane Gabora. Evolutionary approaches to creativity. *The Cambridge Handbook of Creativity*, pages 279–300, 2011.

[20] B. Matthijs, Nijstad, B. A., D. Dreu, and C. K. W. Editorial: "the cognitive, emotional and neural correlates of creativity". *Frontiers in Human Neuroscience*, 9:275, 2015.

[21] G. J. Mills and P. G. T. Healey. Semantic negotiation in dialogue: The mechanisms of alignment. 2008.

[22] M. Minsky. *The Society of Mind*. Simon & Schuster, Inc., New York, NY, USA, 1986.

[23] S. O., H. C. J., , and K. R. Identification and classification of hubs in brain networks. *PLoS ONE*, 2007.

[24] R. Pascanu, T. Weber, S. Racanière, D. P. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, P. Battaglia, D. Silver, and D. Wierstra. Imagination-augmented agents for deep reinforcement learning. *CoRR*, 2017.

[25] J. C. Peterson, P. Soulos, A. Nematzadeh, and T. L. Griffiths. Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *CoRR*, abs/1805.07647, 2018.
[26] H. C. R., J. W. Rieger, M. Cristiano, M. Stephanie, K. R. T., and T. F. E. Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11:61, 2017.
[27] J. RE, S. JM, and B. H. et al. Neuroanatomy of creativity. *Human brain mapping*, 2010.
[28] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *CoRR*, 2008.
[29] I. Singer. *Modes of Creativity: Philosophical Perspectives*. 2013.
[30] P. Singh. Examining the society of mind. *Computing and Informatics*, 22:521–543, 2004.
[31] C. D. T. Blind variation and selective retention in creative thought as in other knowledge processes. *Psychol. Rev. 67*, 1960.
[32] D. Terry. *Creativity, cognition, and knowledge: An interaction*. 2002.
[33] P. Thagard. *Mind: Introduction to cognitive science*. The MIT Press, Cambridge, MA, US, 1996.
[34] G. G. V. and G. D. D. Enhancing user creativity: semantic measures for idea generation. *Knowledge-Based Systems*, 151:1–15, 2018.
[35] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *CoRR*, 2018.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
[37] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *CoRR*, 2018.
[38] R. W. Weisberg. The creative mind versus the creative computer. *Behavioral and Brain Sciences*, 17(3):555–557, 1994.
[39] C. Wong and A. Gesmundo. Transfer learning to learn with multitask neural model search. *CoRR*, 2017.