How Compositional is a Model?

Parikshit Ram, Tim Klinger, Alexander Gray
IBM Research

Context

- **Compositional Generalization is important** in (sequence) learning tasks
- Existing benchmarks **empirically** demonstrate the **lack of compositional generalization** of existing off-the-shelf models
  - With the understanding that these models are not "compositional"

Motivating Questions & Contributions

- What does it mean to be **compositional**?
- Why are some models compositional while others are not?

Motivating Questions & Contributions

- What does it mean to be **compositional**?
- Why are some models compositional while others are not?

- General definition & framework for studying compositional models
- Definition of **compositional complexity**
- Demonstrate how existing models fit this framework, and how they compare against each other

Motivating Questions & Contributions

- How do these definitions of compositionality and compositional complexity relate to guarantees for compositional generalization?

Motivating Questions & Contributions

- How do these definitions of compositionality and compositional complexity relate to guarantees for compositional generalization?

- Demonstrate connection between systematic generalization and compositional complexity

WIP

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

$$\mu(\alpha(u_1, \ldots, u_n)) = r_\alpha(\mu(u_1), \ldots, \mu(u_n))$$

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

meaning function

$$\boxed{\mu}(\alpha(u_1, \ldots, u_n)) = r_\alpha(\mu(u_1), \ldots, \mu(u_n))$$

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

meaning function

$$\underbrace{\mu}_{\text{meaning function}}(\underbrace{\alpha}_{\text{rule}}(\underbrace{u_1, \ldots, u_n}_{\text{sub-terms}})) = r_\alpha(\mu(u_1), \ldots, \mu(u_n))$$

rule **sub-terms**

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

$$\underbrace{\mu}_{\text{meaning function}}(\underbrace{\alpha}_{\text{rule}}(\underbrace{u_1, \ldots, u_n}_{\text{sub-terms}})) = \underbrace{r_\alpha}_{\text{rule-dep function}}(\mu(u_1), \ldots, \mu(u_n))$$

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

$$\underbrace{\mu}_{\text{meaning function}}(\underbrace{\alpha}_{\text{rule}}(\underbrace{u_1, \ldots, u_n}_{\text{sub-terms}})) = \underbrace{r_\alpha}_{\text{rule-dep function}}(\underbrace{\mu(u_1), \ldots, \mu(u_n)}_{\text{meanings of the sub-terms}})$$

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

$$\underbrace{\mu}_{\text{meaning function}}(\underbrace{\alpha}_{\text{rule}}(\underbrace{u_1, \ldots, u_n}_{\text{sub-terms}})) = \underbrace{r_\alpha}_{\text{rule-dep function}}(\underbrace{\mu(u_1), \ldots, \mu(u_n)}_{\text{meanings of the sub-terms}})$$

Recursive

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.

Defining Compositionality

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

meaning function

$$\underbrace{\mu}_{\text{meaning function}}(\underbrace{\alpha}_{\text{rule}}(\underbrace{u_1, \ldots, u_n}_{\text{sub-terms}})) = \underbrace{r_\alpha}_{\text{rule-dep function}}(\underbrace{\mu(u_1), \ldots, \mu(u_n)}_{\text{meanings of the sub-terms}})$$

rule sub-terms rule-dep function meanings of the sub-terms

Recursive

Any meaning function can be shown to have compositional semantics

Partee, Barbara. "Lexical semantics and compositionality." *An invitation to cognitive science: Language* 1 (1995): 311-360.
Pagin, Peter, and Dag Westerståhl. "Compositionality I: Definitions and variants." *Philosophy Compass* 5.3 (2010): 250-264.
Zadrozny, Wlodek. "From compositional to systematic semantics." *Linguistics and philosophy* 17 (1994): 329-342.

Our Proposed Definition: Components

Components of a compositional model for sequences:

Our Proposed Definition: Components

Components of a compositional model for sequences:

- *Token encoder*: Encodes input tokens in latent space $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
    - Can handle positional encoding

Our Proposed Definition: Components

Components of a compositional model for sequences:

- *Token encoder*: Encodes input tokens in latent space $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
  - Can handle positional encoding
- *Span processor*: Processes sequence spans $g : \mathcal{H}^k \to \mathcal{H}$
  - Can itself be a collection of functions

Our Proposed Definition: Components

Components of a compositional model for sequences:

- *Token encoder*: Encodes input tokens in latent space $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
  - Can handle positional encoding
- *Span processor*: Processes sequence spans $g : \mathcal{H}^k \to \mathcal{H}$
  - Can itself be a collection of functions
- *Computational DAG function*: The processing hierarchy $D : \mathcal{X} \to \mathcal{D}$ *(the space of DAGs),*
  - The trace of a **program** processing the sequence
  - Can be input-dependent

Our Proposed Definition: Components

Components of a compositional model for sequences:

- *Token encoder*: Encodes input tokens in latent space $e : \mathcal{I} \times \mathbb{N} \rightarrow \mathcal{H}$
  - Can handle positional encoding
- *Span processor*: Processes sequence spans $g : \mathcal{H}^k \rightarrow \mathcal{H}$
  - Can itself be a collection of functions
- *Computational DAG function*: The processing hierarchy $D : \mathcal{X} \rightarrow \mathcal{D}$ *(the space of DAGs)*
  - The trace of a **program** processing the sequence
  - Can be input-dependent
- *Read-out function*: Outputs the final "meaning" $h : \mathcal{H}^m \rightarrow \mathcal{Y}$
  - The label can be class, target or next-token

Our Proposed Definition: Compositional Function

Components of a compositional model for sequences:

- *Token encoder* $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
- *Span encoder* $g : \mathcal{H}^k \to \mathcal{H}$
- *Computational DAG function* $D : \mathcal{X} \to \mathcal{D}$ *(the space of DAGs)*
- *Read-out function* $h : \mathcal{H}^m \to \mathcal{Y}$

Our Proposed Definition: Compositional Function

Components of a compositional model for sequences:

- *Token encoder* $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
- *Span encoder* $g : \mathcal{H}^k \to \mathcal{H}$
- *Computational DAG function* $D : \mathcal{X} \to \mathcal{D}$ *(the space of DAGs)*.
- *Read-out function* $h : \mathcal{H}^m \to \mathcal{Y}$

Compositional function: $f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$

Our Proposed Definition: Compositional Function

Components of a compositional model for sequences:

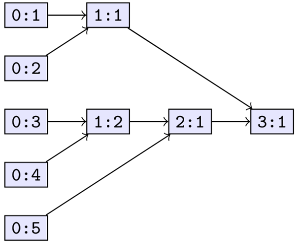
- *Token encoder* $e : \mathcal{I} \times \mathbb{N} \rightarrow \mathcal{H}$
- *Span encoder* $g : \mathcal{H}^k \rightarrow \mathcal{H}$
- *Computational DAG function* $D : \mathcal{X} \rightarrow \mathcal{D}$ *(the space of DAGs)*
- *Read-out function* $h : \mathcal{H}^m \rightarrow \mathcal{Y}$

Compositional function: $f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$

Recursive application of the span encoder over the computational DAG

Compositional Model: Example

$$f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$$



CDAG ex #1

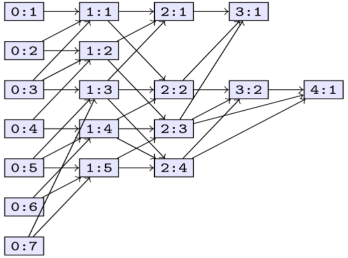
$$X = [x_1, \ldots, x_5]$$

$$e_i = e(x_i, i) \in \mathcal{H}$$

$$f(X) = h\left(g\left(e_1, e_2\right), g\left(g(e_3, e_4), e_5\right)\right)$$

Compositional Model: Another Example

$$f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$$



CDAG ex #2

$$X = [x_1, \ldots, x_7]$$
$$e_i = e(x_i, i) \in \mathcal{H}$$
$$f(X) = h\left(v_{4:1}, v_{3:1}\right)$$

$$v_{0:i} = e_i$$
$$v_{1:1} \leftarrow g(e_1, e_2, e_3)$$
$$v_{2:3} \leftarrow g(v_{1:2}, v_{1:4}, v_{1:5})$$
$$v_{3:2} \leftarrow g(v_{2:2}, v_{2:3}, v_{2:4})$$
$$v_{4:1} \leftarrow g(v_{3:2}, v_{2:3}, v_{2:4})$$

Compositional Models

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

Compositional Models

The meaning of the whole is a function of the meaning of the parts and of the way they are syntactically combined.

$$\underbrace{\text{The meaning of the whole}}_{f:\mathcal{X}\to\mathcal{Y}} \text{ is a } \underbrace{\text{function}}_{h:\mathcal{H}^m\to\mathcal{Y}} \text{ of } \underbrace{\text{the meanings of the parts}}_{g:\mathcal{H}^k\to\mathcal{H}}$$

$$\text{and } \underbrace{\text{of the way they are syntactically combined.}}_{D:\mathcal{X}\to\mathcal{D}}$$

Our Proposed Definition: Compositional Function

Components of a compositional model for sequences:

- *Token encoder* $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
- *Span encoder* $g : \mathcal{H}^k \to \mathcal{H}$
- *Computational DAG function* $D : \mathcal{X} \to \mathcal{D}$ *(the space of DAGs)*
- *Read-out function* $h : \mathcal{H}^m \to \mathcal{Y}$

Compositional function: $f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$

Our Proposed Definition: Compositional Function

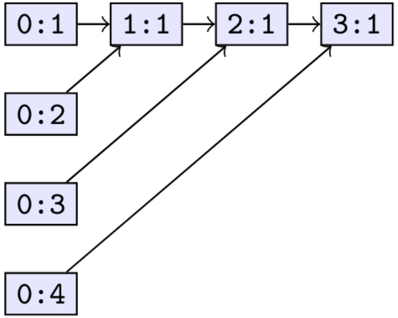
Components of a compositional model for sequences:

- *Token encoder* $e : \mathcal{I} \times \mathbb{N} \to \mathcal{H}$
- *Span encoder* $g : \mathcal{H}^k \to \mathcal{H}$
- *Computational DAG function* $D : \mathcal{X} \to \mathcal{D}$ *(the space of DAGs)*
- *Read-out function* $h : \mathcal{H}^m \to \mathcal{Y}$

Compositional function: $f(X) = h\left(g^{\otimes D(X)}(e(x_1, 1), \ldots, e(x_L, L))\right)$

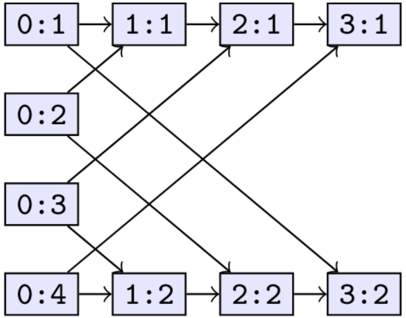
Is this an interesting class of models?

Recurrent Models fit this Framework



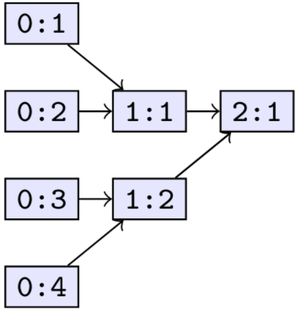
Unidirectional-RNN

$$h(g(g(g(e_1, e_2), e_3), e_4))$$



Bidirectional-RNN

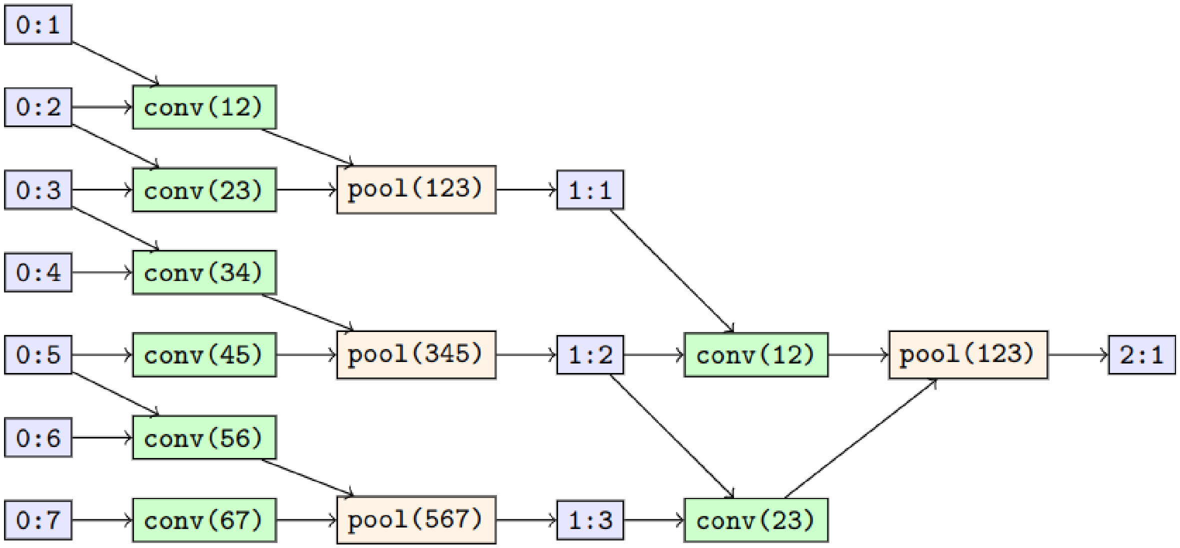
$$h(g(g(g(e_1, e_2), e_3), e_4), g(g(g(e_4, e_3), e_2), e_1))$$



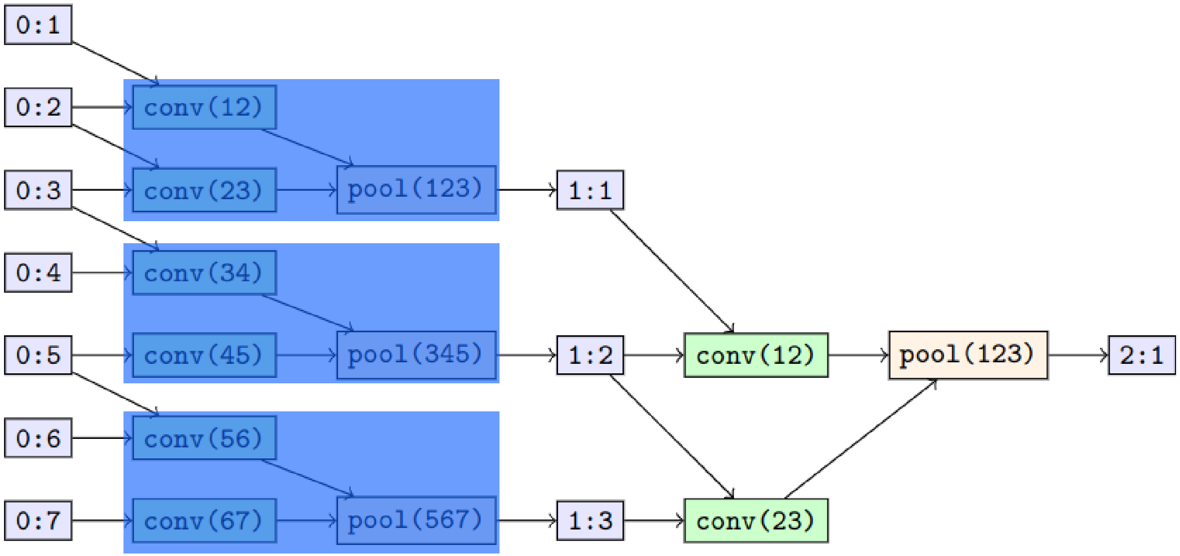
Tree-RNN

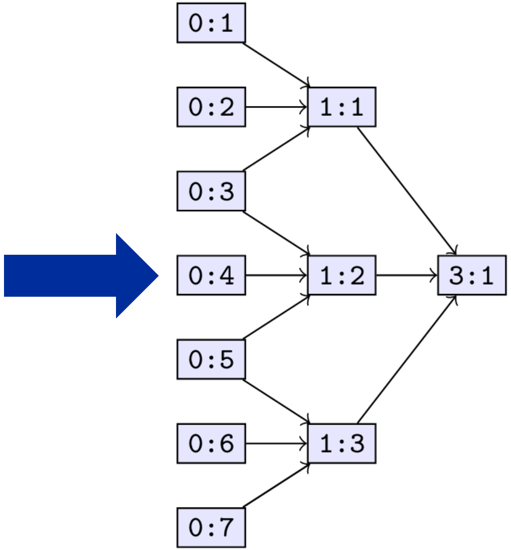
$$h(g(g(e_1, e_2), g(e_3, e_4)))$$

Convolutional Models fit this Framework

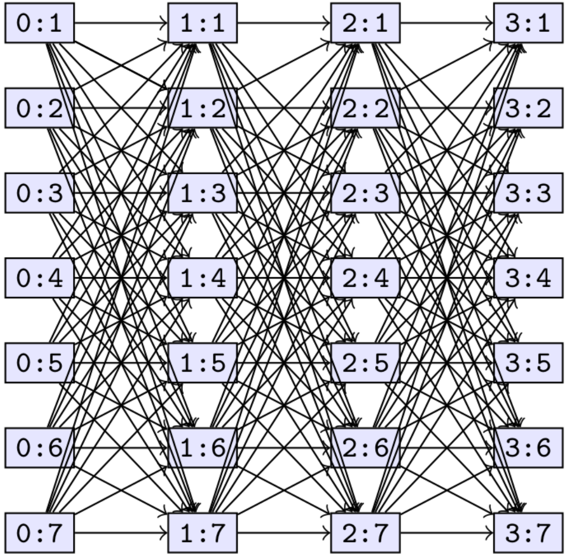


Convolutional Models fit this Framework



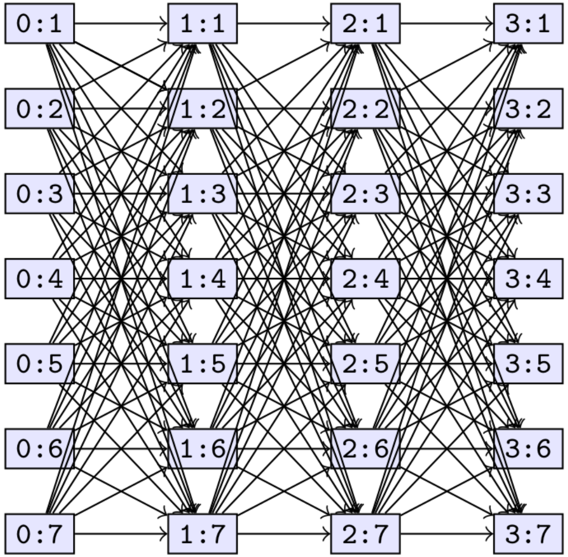


Multi-layered Models fit this Framework

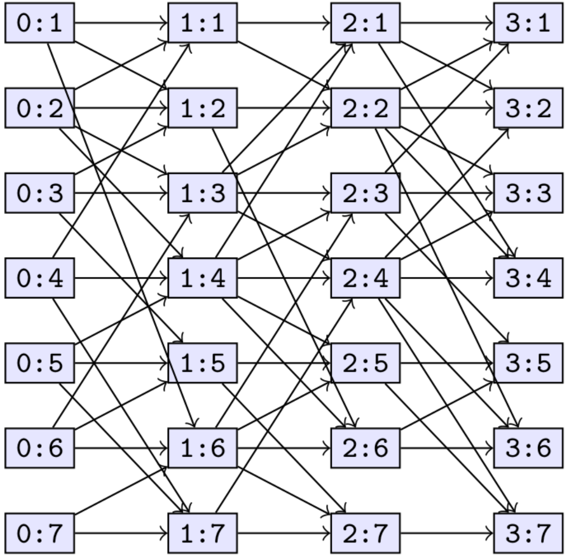


Multi-layered Fully-Connected

Multi-layered Models fit this Framework



Multi-layered Fully-Connected



Multi-layered Fully-Connected w/ sparse or hard top-K attention

Model Comparison

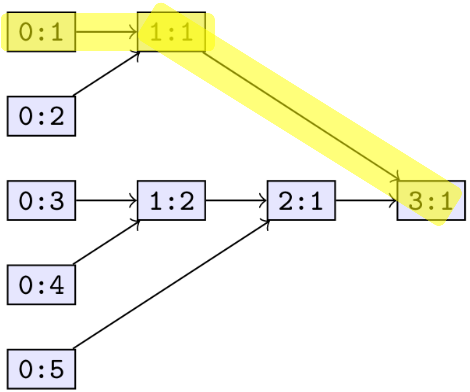
Compositional Model	Arbitrary Length Operation	Input-dependent cDAG
Unidirectional recurrence	✓	✗
Bidirectional recurrence	✓	✗
Tree recurrence	✓	✗
Convolution-then-pooling	✓	✓ **
Multi-layered fully-connected	✗	✗
Multi-layered FC w/ sparse / hard attention	✗	✓

Compositional Complexity

Locus of Influence or LoI of source node

- Encodes complexity of span encoder & cDAG
- Quantifies sensitivity of function output to changes in specific input tokens
- Absolute LoI measures absolute sensitivity
- Relative LoI measures how sensitive a source node is relative to other source nodes

Compositional Complexity



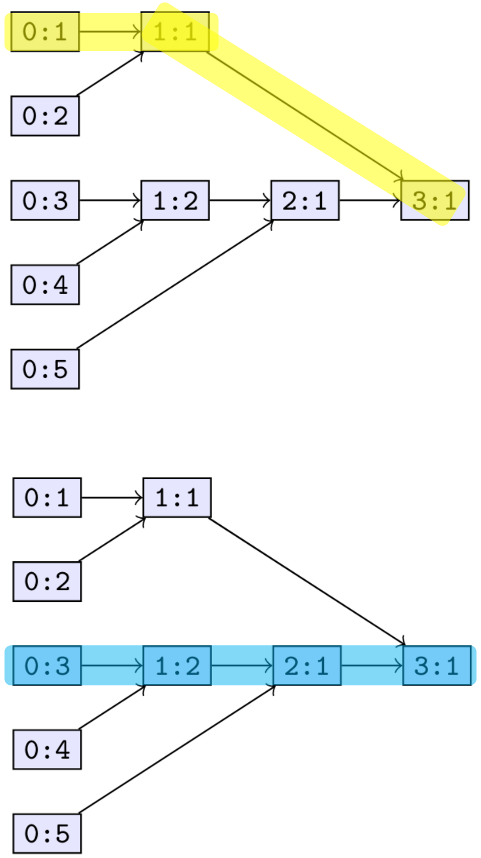
Locus of Influence or LoI of source node

- Encodes complexity of span encoder & cDAG
- Quantifies sensitivity of function output to changes in specific input tokens
- Absolute LoI measures absolute sensitivity
- Relative LoI measures how sensitive a source node is relative to other source nodes

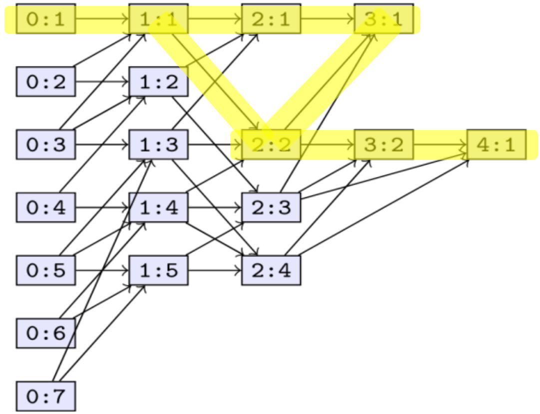
Compositional Complexity

Locus of Influence or LoI of source node

- Encodes complexity of span encoder & cDAG
- Quantifies sensitivity of function output to changes in specific input tokens
- Absolute LoI measures absolute sensitivity
- Relative LoI measures how sensitive a source node is relative to other source nodes



Compositional Complexity



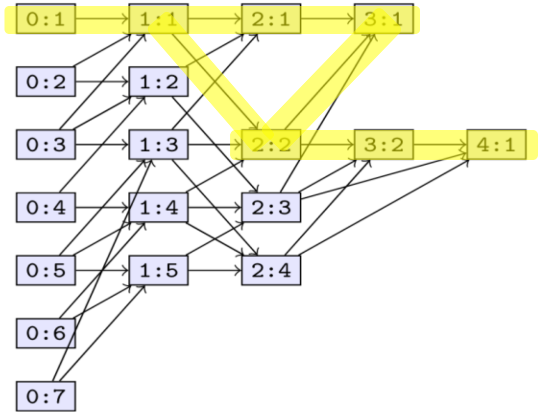
Locus of Influence or LoI of source node

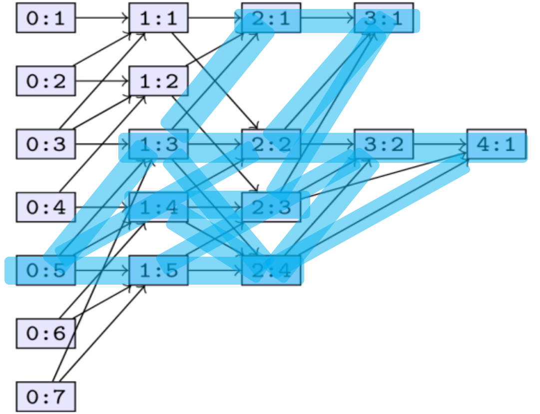
- Encodes complexity of span encoder & cDAG
- Quantifies sensitivity of function output to changes in specific input tokens
- Absolute LoI measures absolute sensitivity
- Relative LoI measures how sensitive a source node is relative to other source nodes

Compositional Complexity

Locus of Influence or LoI of source node

- Encodes complexity of span encoder & cDAG
- Quantifies sensitivity of function output to changes in specific input tokens
- Absolute LoI measures absolute sensitivity
- Relative LoI measures how sensitive a source node is relative to other source nodes





Compositional Function Class

Class of Compositional Functions: Functions in this class have a bounded absolute & relative LoI for any input sequence and corresponding source nodes in the cDAG

- Small bounds on absolute & relative LoI imply **simple compositional functions**
- Large absolute LoI & small relative LoI imply **extremely complex compositional functions**
- (Moderately) large absolute LoI & large relative LoI imply functions allow **some tokens** in the input sequence to have **a lot of influence**, but **most tokens** have **simple compositional structure**

Model Comparison

Model	Absolute LoI	Relative LoI
Unidirectional recurrence	c^{L-1}	$1/2$
Bidirectional recurrence	c^{L-1}	$1/4$
Tree recurrence	$c^{\log L}$	$1/L$
Convolution-then-pooling	$c^{\log L}$	$\frac{2}{L(1+\frac{1}{p})}$
Multi-layered fully-connected	$(Lc)^M$	$1/L$
Multi-layered fully-connected w/ sparse/hard attention	$L(Kc)^M$	$1/K$

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

$$\Delta \triangleq \max_{\substack{D,g,h,\\ f:=\{e,D,g,h\},\\ f\in\mathcal{F},\\ X\in\mathcal{X}}} \min_{\substack{\hat{\mathbf{D}},\hat{g},\hat{h},\\ \hat{f}:=\{e,\hat{\mathbf{D}},\hat{g},\hat{h}\},\\ \hat{f}\in\mathcal{F}}} \left| h(g^{\otimes D(X)}(e(x_1),\ldots,e(x_L)) - \hat{h}(\hat{g}^{\otimes \hat{\mathbf{D}}}(e(x_1),\ldots,e(x_L)) \right|$$

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

$$\Delta \triangleq \max_{\substack{D,g,h,\\ f:=\{e,D,g,h\},\\ f\in\mathcal{F},\\ X\in\mathcal{X}}} \min_{\substack{\hat{\mathbf{D}},\hat{g},\hat{h},\\ \hat{f}:=\{e,\hat{\mathbf{D}},\hat{g},\hat{h}\},\\ \hat{f}\in\mathcal{F}}} \left| h(g^{\otimes D(X)}(e(x_1),\ldots,e(x_L)) - \hat{h}(\hat{g}^{\otimes \hat{\mathbf{D}}}(e(x_1),\ldots,e(x_L)) \right|$$

approximation

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

$$\boxed{\Delta} \triangleq \max_{\substack{D, g, h, \\ f := \{e, D, g, h\}, \\ f \in \mathcal{F}, \\ X \in \mathcal{X}}} \min_{\substack{\hat{\mathrm{D}}, \hat{g}, \hat{h}, \\ \hat{f} := \{e, \hat{\mathrm{D}}, \hat{g}, \hat{h}\}, \\ \hat{f} \in \mathcal{F}}} \left| h(g^{\otimes D(X)}(e(x_1), \ldots, e(x_L)) - \hat{h}(\hat{g}^{\otimes \hat{\mathrm{D}}}(e(x_1), \ldots, e(x_L)) \right|$$

approx

functions of same complexity

input-dependent cDAG

input-agnostic cDAG

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

$$\Delta \triangleq \max_{\substack{D,g,h, \\ f:=\{e,D,g,h\}, \\ f\in\mathcal{F}, \\ X\in\mathcal{X}}} \min_{\substack{\hat{\mathbf{D}},\hat{g},\hat{h}, \\ \hat{f}:=\{e,\hat{\mathbf{D}},\hat{g},\hat{h}\}, \\ \hat{f}\in\mathcal{F}}} \left| h(g^{\otimes D(X)}(e(x_1),\ldots,e(x_L)) - \hat{h}(\hat{g}^{\otimes \hat{\mathbf{D}}}(e(x_1),\ldots,e(x_L))\right|$$

$$C_l \delta \leq \Delta \leq C_u \frac{\delta}{\beta}$$

Absolute LoI

Relative LoI

Impact of Input-dependent cDAG

How well can a compositional function with an input-agnostic cDAG approximate a compositional function (of same complexity) with input-dependent cDAGs?

$$C_l \delta \leq \Delta \leq C_u \frac{\delta}{\beta}$$

Absolute LoI

Relative LoI

Input-agnostic cDAG cannot sufficiently approximate input-dependent cDAGs

- More complex the composition, the worse the approximation
- Lower absolute LoI allows for better approximation
- For same absolute LoI, higher relative LoI allows for better approximation
- Models with input-dependent cDAGs can be more expressive than models with input-agnostic cDAGs

Quantifying Systematic Generalization

Systematic generalization is often described as being able to "*handle unknown combination of known parts*"

Quantifying Systematic Generalization

WIP

Systematic generalization is often described as being able to "*handle unknown combination of known parts*"

Learning setup:

- Given token encoder & cDAG function, learn the span encoder & the readout function
- Considering "exchangeable parts": Subsequences $X, V \in \mathcal{I}^*$

  such that $X_1 X X_2 \in \mathcal{X}$ and $X_1 V X_2 \in \mathcal{X}$ for some prefix/suffix $X_1, X_2 \in \mathcal{I}^*$

- Define (implicitly) known combinations of (implicitly) known parts via low error on examples

Ground-truth $$\left| \hat{h} \circ \hat{g}^{\otimes D(X_1 X X_2)}(e(X_1 X X_2)) - h \circ g^{\otimes D(X_1 X X_2)}(e(X_1 X X_2)) \right| \leq \epsilon$$

Learned $$\left| \hat{h} \circ \hat{g}^{\otimes D(V_1 V V_2)}(e(V_1 V V_2)) - h \circ g^{\otimes D(V_1 V V_2)}(e(V_1 V V_2)) \right| \leq \epsilon$$

Quantifying Systematic Generalization

WIP

Systematic generalization is often described as being able to "*handle unknown combination of known parts*"

Learning setup:

- Unknown combination of known parts: $X_1 V X_2$
- What can we say about the quality of the learned functions on the unknown combination

$$\left| \hat{h} \circ \hat{g}^{\otimes D(X_1 V X_2)}(e(X_1 V X_2)) - h \circ g^{\otimes D(X_1 V X_2)}(e(X_1 V X_2)) \right|$$

- Preliminary results indicate that more complex compositional functions will have a worse bound
- Learning complex compositions is hard even if the cDAG is given and we just need to learn the span encoder and readout function

Thank you

Parikshit Ram
Parikshit.Ram@ibm.com