

What they do when in doubt: a study of inductive bias in seq2seq learners

Eugene Kharitonov, Rhama Chaabouni

紹介者: 栗林樹生 (東北大)

※この研究の解説というよりは、とりまく背景・展望について

私達は今異世界に転生しました。



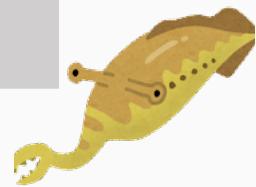
<https://item.rakuten.co.jp/to-ki/4589588274127/>

私達は今異世界に転生しました。

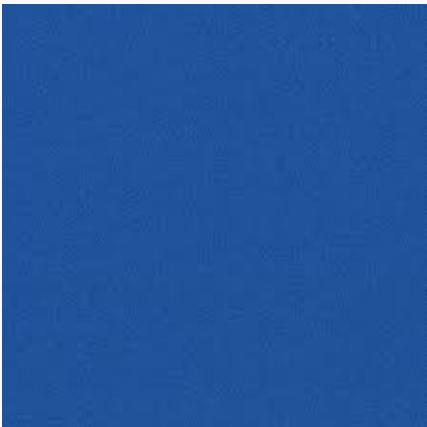


<https://item.rakuten.co.jp/to-ki/4589588274127/>

これはボコトっていうんだ



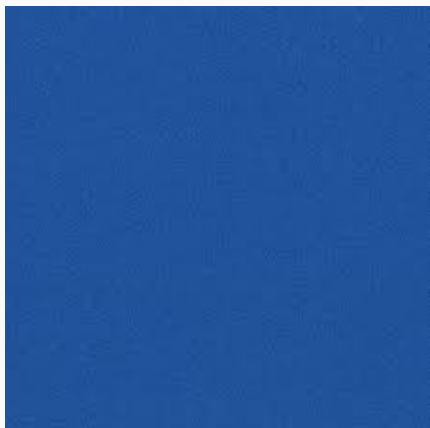
どれがボコトですか？



www.amazon.co.jp/dp/B00IK5URXM

LA' vie avec' Brocante.

どれがボコトですか？



これがボコトかな？



www.amazon.co.jp/dp/B00IK5URXM

LA' vie avec' Brocante.

同じ形をしたものに同じ名前が割り当てられていると
一般化する傾向 Shape bias [Landau et al., 1988]



ちなみに、ニューラルモデルも
shape-biasをもつらしい

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study [Ritter+, 2017]
Learning Inductive Biases with Simple Neural Networks [Feinman & Lake, 2018]

aabaa → *b*

bbabb → *a*

aba → ?

$aabaa \rightarrow b$
 $bbabb \rightarrow a$

$aba \rightarrow$

Linear learner:

a

3文字目を出力していた

$aabaa \rightarrow b$
 $bbabb \rightarrow a$
③

Hierarchical learner:

b

ネストの中を出力していた

$aabaa \rightarrow b$
 $bbabb \rightarrow a$

子供は、言語習得においてhierarchicalな一般化を好む
[Perfors+, 2001]



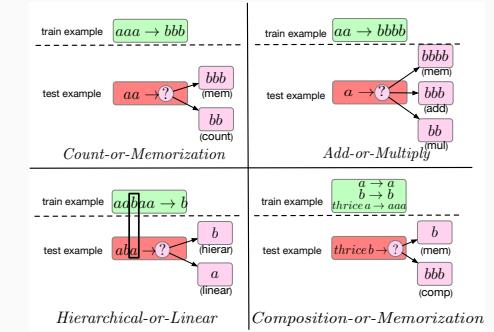
ちなみにCNNはLinear派、Transformer, LSTMは
hierarchical派らしい [Kharitonov&Chaabouni, 2021]

概要: ニューラルネットの帰納バイアス

- ニューラルモデルの帰納バイアス (考え方の癖) を分析
 - ごく少数(单一)のインスタンスからどのような汎化をするか

train $aabaa \rightarrow b$
 $bbabb \rightarrow a$

test $aba \rightarrow ?$



- モデルアーキテクチャ (LSTM, CNN, Transformer) とドロップアウト率が特にモデルの帰納バイアスに影響を与える
 - 層数, 隠れ層の次元数, 埋め込み層の次元数, オプティマイザ (SGD or Adam) は微小な影響
 - (residualも関連しそうだが、触れられていない)

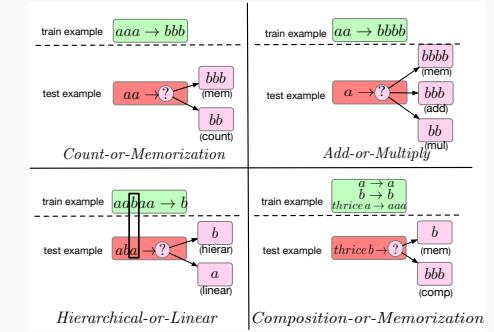
概要: ニューラルネットの帰納バイアス

- ニューラルモデルの帰納バイアス (考え方の癖) を分析

- ごく少数(单一)のインスタンスからどのような汎化をするか

train $\begin{array}{ll}aabaa & \rightarrow b \\ bbabb & \rightarrow a\end{array}$

test $aba \rightarrow ?$



- モデルアーキテクチャ (LSTM, CNN, Transformer) とドロップアウト率が特にモデルの帰納バイアスに影響を与える

- 層数, 隠れ層の次元数, 埋め込み層の次元数, オプティマイザ (SGD or Adam) は微小な影響
- (residualも関連しそうだが, 触れていない)

- だから何なのよ? (by reviewer 5)

今日の話の主題

On the proper role of linguistically-oriented deep net analysis in linguistic theorizing [Baroni, 2021] が代弁



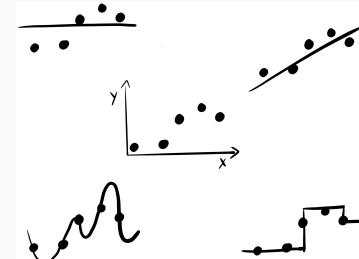
Deep netsが言語の諸側面を捉えてい
ることが、言語に
ついて何を語って
いることになるの
か？(の一考)

帰納バイアス (機械学習)

● Inductive bias

- **The set of assumptions** that the learner uses **to predict outputs of given inputs that it has not encountered**.
- **An unbiased learning system's** ability to classify new instances is no better than if it **simply stored** all the training instances and performed a lookup when asked to classify a subsequent instance [Mitchell 1980]

```
class MyModel:  
    def __init__(self):  
        self.memory = dict()  
  
    def train(self, data):  
        for x, y in data:  
            self.memory[x] = y  
  
    def infer(self, x):  
        try:  
            return self.memory[x]  
        except:  
            raise UnseenError("I've never seen data like this before, \  
so why could I make a prediction?")
```



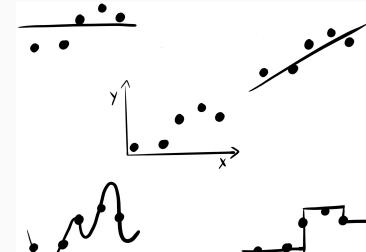
<https://www.mlnar.com/inductive-bias.html>

帰納バイアス (機械学習)

● Inductive bias

- **The set of assumptions** that the learner uses **to predict outputs of given inputs that it has not encountered**.
- **An unbiased learning system's** ability to classify new instances is no better than if it **simply stored** all the training instances and performed a lookup when asked to classify a subsequent instance [Mitchell 1980]

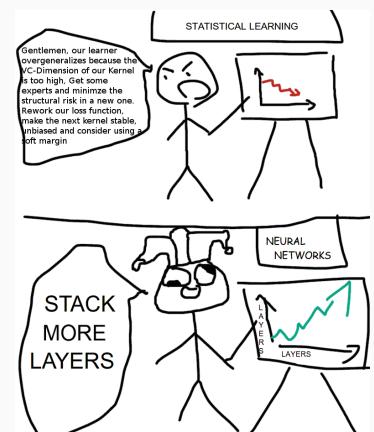
```
class MyModel:  
    def __init__(self):  
        self.memory = dict()  
  
    def train(self, data):  
        for x, y in data:  
            self.memory[x] = y  
  
    def infer(self, x):  
        try:  
            return self.memory[x]  
        except:  
            raise UnseenError("I've never seen data like this before, \  
so why could I make a prediction?")
```



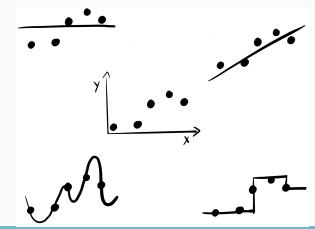
<https://www.mlnar.com/inductive-bias.html>

● 有限のデータから未知のデータについて何か言う際には仮定 (帰納バイアス) が必要

- 関係が線形だと思おう. 入力が近ければ出力も近くなるだろう.
ニューラルネットでsoftな表現にしよう. Transformer layerを積みまくろう...

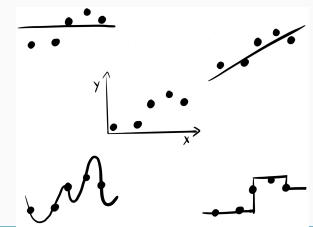


<https://stackmorelayers.be/>



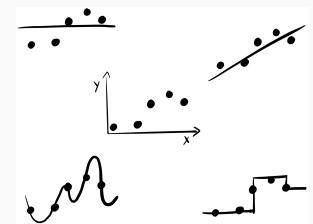
人の帰納バイアス (言語獲得)

- 私たちは有限個の事例_(文)から、見たことのない文を書いたり読んだりできるようになる。しかも、皆同じ文法を獲得する (刺激の貧困, Chomsky)



人の帰納バイアス (言語獲得)

- 私たちは有限個の事例_(文)から、見たことのない文を書いたり読んだりできるようになる。しかも、皆同じ文法を獲得する (刺激の貧困, Chomsky)
- 何らかの帰納バイアスに従って、言語獲得が進んでいるように見える
 - それはどんなバイアスか？
 - 人は**生得的に何かを持っているのか？**
 - 人間の本質はなにか？



人の帰納バイアス (言語獲得)

- 私たちは有限個の事例(文)から、見たことのない文を書いたり読んだりできるようになる。しかも、皆同じ文法を獲得する (刺激の貧困, Chomsky)
- 何らかの帰納バイアスに従って、言語獲得が進んでいるように見える
 - それはどんなバイアスか？
 - 人は**生得的に何かを持っているのか？**
 - 人間の本質はなにか？

We do not learn, and that what we call learning is only a process of recollection.
-- Plato

All men are created equal
-- Thomas Jefferson

I think; therefore I am
-- Rene Descartes

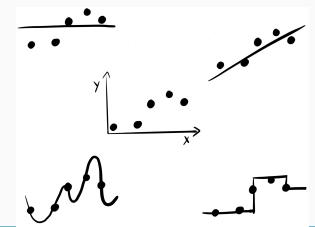
A consistent materialist would consider it as self-evident that the mind has very important innate structures, physically realized in some manner. Why should it be otherwise?

-- Noam Chomsky

No man's knowledge here can go beyond his experience.
-- John Locke

There has to be innate circuitry that does the learning, that creates the culture, that acquires the culture, and that responds to socialization.
-- Steven Pinker [スティーブン・ピンカー：書き込まれた「空白の石版」 - TED](#)

長きに亘る nature-nurture debate



人の帰納バイアス (言語獲得)

- 私たちは有限個の事例(文)から、見たことのない文を書いたり読んだりできるようになる。しかも、皆同じ文法を獲得する (刺激の貧困, Chomsky)
- 何らかの帰納バイアスに従って、言語獲得が進んでいるように見える
 - それはどんなバイアスか？
 - 人は**生得的に何かを持っているのか？**
 - 人間の本質はなにか？

We do not learn, and that what we call
learning is only a process of recollection.
-- Plato

All men are created equal
-- Thomas Jefferson

I think; therefore I am
-- Rene Descartes

A consistent materialist would consider it as self-evident that the mind has very
important innate structures, physically realized in some manner. Why should it be
otherwise?

-- Noam Chomsky

No man's knowledge here can go beyond his experience.
-- John Locke

There has to be innate circuitry that does the learning, that creates the
culture, that acquires the culture, and that responds to socialization.
-- Steven Pinker [スティーブン・ピンカー：書き込まれた「空白の石版」 - TED](#)

Artificial neural networks learn some kinds of linguistic structure only from data !!!???
-- 21st natural language processing

言語は経験的に獲得できるということ？？
今日の紹介で少しでも洞察を与えられたらOK

Emergent linguistic structure in artificial neural
networks trained by self-supervision

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy
+ See all authors and affiliations

人の言語獲得に関する帰納バイアスをどう調べるか

この分け方は [McCoy+, 2018] より

- 人を調べる

- 子供が言語の階層的な一般化を好むことを心理実験などで示す [Perfors+, 2001] など

人の言語獲得に関する帰納バイアスをどう調べるか

この分け方は [McCoy+, 2018] より

- 人を調べる
 - 子供が言語の階層的な一般化を好むことを心理実験などで示す [Perfors+, 2001]など
- 数理モデルを用いる(構成論的なアプローチ)
 - 議論の例: 何らかのモデル(人らしい生得的能力を持っていないままさらな状態と仮定)が、データのみからX(例. 文法)を獲得できるならば、Xは経験的に獲得できる(Xを生得的に獲得していることが言語獲得のための必要条件ではない)
 - "The APS [(argument from the poverty of the stimulus)] predicts that any artificial learner trained with no prior knowledge of the principles of syntax [...] must fail to make acceptability judgments with human-level accuracy.[...] If linguistically uninformed neural network models achieve human-level performance on specific phenomena [...], this would be clear evidence limiting the scope of phenomena for which the APS can hold" [Warstadt+, 2019]
 - "Our results also contribute to the long-running nature-nurture debate in language acquisition: whether the success of neural models implies that unbiased learners can learn natural languages with enough data, or whether human abilities to acquire language given sparse stimulus implies a strong innate human learning bias" [Papadimitriou & Jurafsky, 2020]
 - [Baroni, 2021]より引用、このような論法を**blank slate argument**と呼んでいる。

人の言語獲得に関する帰納バイアスをどう調べるか

この分け方は [McCoy+, 2018] より

- 人を調べる

- 子供が言語の階層的な一般化を好むことを心理実験などで示す [Perfors+, 2001]など

- 数理モデルを用いる(構成論的なアプローチ)

妥当？

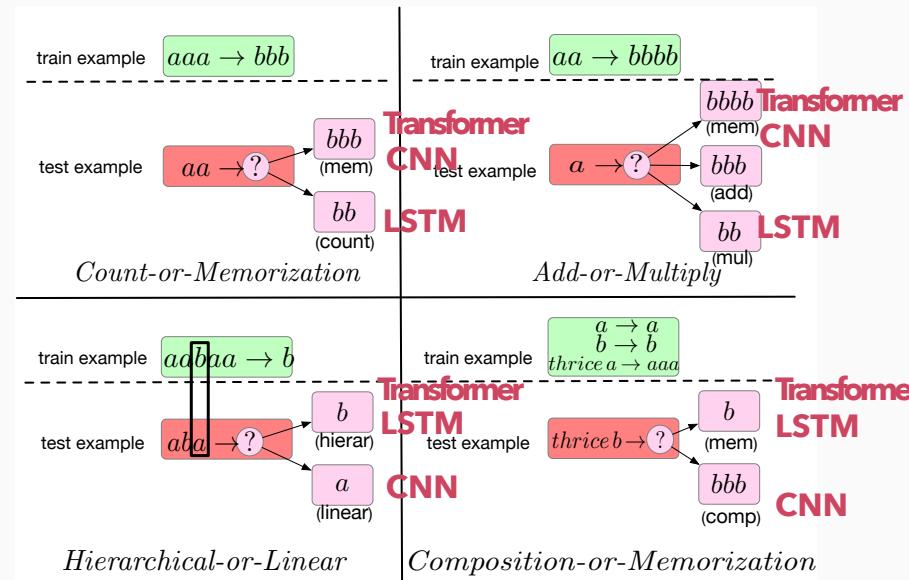
- 議論の例: 何らかのモデル (**人らしい生得的能力を持っていないまっさらな状態と仮定**) が、データのみからX (例. 文法) を獲得できるならば、Xは経験的に獲得できる (Xを生得的に獲得していることが言語獲得のための必要条件ではない)

- "The APS [(argument from the poverty of the stimulus)] predicts that any artificial learner trained with no prior knowledge of the principles of syntax [...] must fail to make acceptability judgments with human-level accuracy.[...] If linguistically uninformed neural network models achieve human-level performance on specific phenomena [...], this would be clear evidence limiting the scope of phenomena for which the APS can hold" [Warstadt+, 2019]
- "Our results also contribute to the long-running nature-nurture debate in language acquisition: whether the success of neural models implies that unbiased learners can learn natural languages with enough data, or whether human abilities to acquire language given sparse stimulus implies a strong innate human learning bias" [Papadimitriou & Jurafsky, 2020]
- [Baroni, 2021]より引用、このような論法を**blank slate argument**と呼んでいる。

こういうことを言っていいのか？
短絡的では？ [Baroni, 2021]

再掲: ニューラルネットの帰納バイアス

- ニューラルモデルがごく少数(单一)のインスタンスからどのような汎化を好むかを通して、モデルの帰納バイアス(initの時点で獲得した考え方の癖)を分析



- 人らしい帰納バイアスを持っていないという仮定は強すぎるので、**blank slate argument**は短絡的
 - モデルによっては、階層的な汎化を好むなど人らしい帰納バイアスを既に有している

テクニカルな面: 記述長に基づいた評価

$aa \rightarrow bbbb$

$a^k \rightarrow b^{2k}$

出力が常に $bbbb$

- あるモデル M が、学習データ $T = \{x_1, y_1\}$ から、規則 A, B どちらの汎化を好むかを知りたい
- 気持ち:
モデル M を学習事例 T で事前訓練した後、ある規則に従うデータ D で追加学習する。
validation loss の収束が早い \Leftrightarrow その規則を好む帰納バイアスを有する

簡略化のため
に結構嘘を
ついています。
気になる人は
論文を読んで
欲しい

面白いけど割愛 (不必要に実験手順を複雑化させてかっこよく見せているように見える。 Appendix)

情報理論との接続

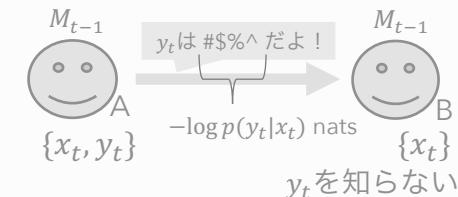
具体的に計算しているものは結構トリッキー

各step t にて Adam で 300 epoch 回す、
optimizer は reset しない

$$L_M(D) = - \sum_{t=2}^k \log p_{M_{t-1}}(y_t | x_t) + c.$$

M_1 は学習データで学習したモデル。 M_{t-1} は、データ $\{x_i, y_i\}_{i=1}^{t-1}$ で学習したモデル。
 D からデータを 1 つ取り出す → そのデータの $-\log p(y_t | x_t)$ を計算する
→ 学習データに加えて M_t を訓練する (from scratch) を繰り返す。

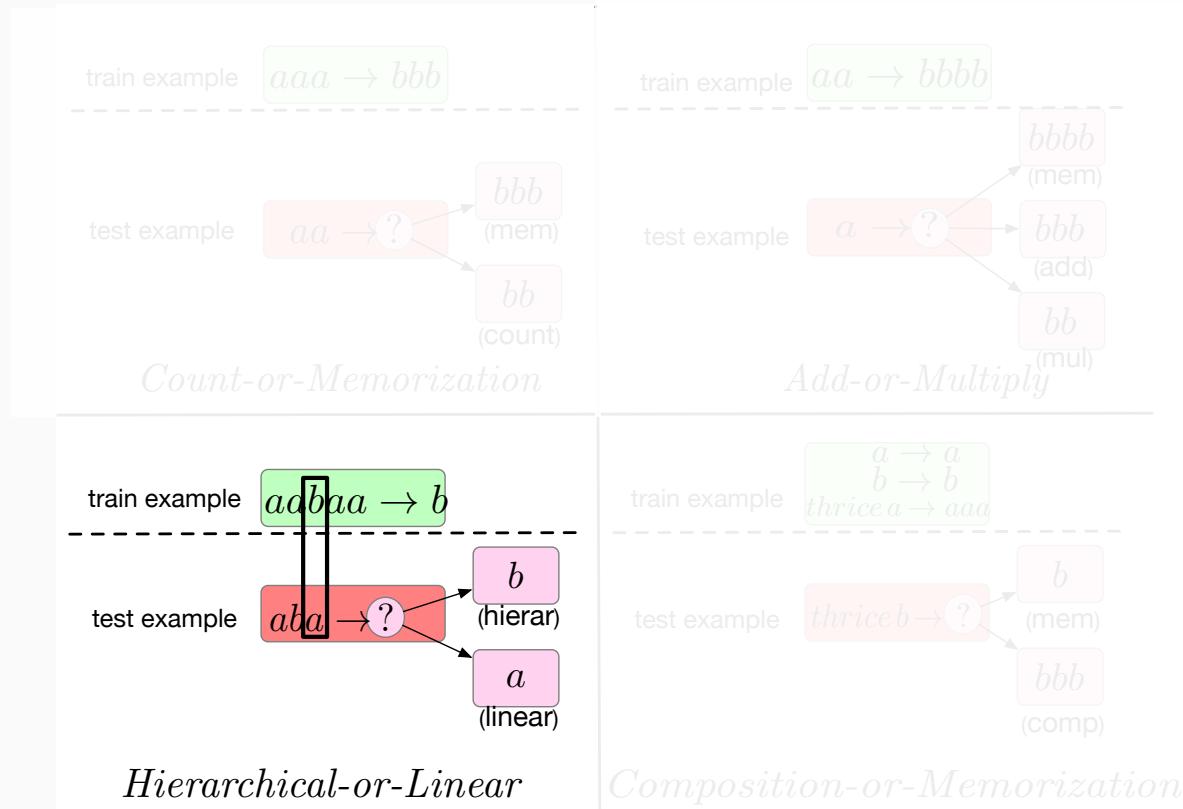
- 学習データで学習したモデルにとって、どちらの規則に従うデータ D_A, D_B が「簡潔か」を調べる
 - この簡潔さとして、情報理論の言葉でモデルとデータの圧縮転送コスト (記述長) を考える (prequential codes) [Dawid, 1984] [Blier&Olliver, 2018]



$\{x_t, y_t\}$ を知っている A が $\{x_t\}$ しか知らない B に、 $\{y_t\}$ と最終的にそのデータで学習されるモデルのパラメータを転送したい。どのモデルをどのような設定でアップデートするかはお互い合意がされている。データ $\{y_t\}$ をブロックに分けて圧縮して転送し、転送済みのデータで両者のモデルをアップデートするという手続きを繰り返すことで転送を行う (prequential codes)。このとき転送にかかった総コスト (nats) を、そのデータとモデルの記述長とみなす

ニューラルネットの帰納バイアス

● 検証例1: Hierarchical or Linear



小さい: その規則を好む

	FPA ↑ hierar	FPA ↑ linear	L , nats ↓ hierar	L , nats ↓ linear
LSTM-s2s no att.	0.05	0.00	31.04*	61.84
LSTM-s2s att.	0.30	0.00	26.32*	57.2
CNN-s2s	0.00	1.00	202.64	0.00*
Transformer	0.69	0.00	4.84*	35.04

(c) Hierarchical-or-Linear with depth $d = 4$

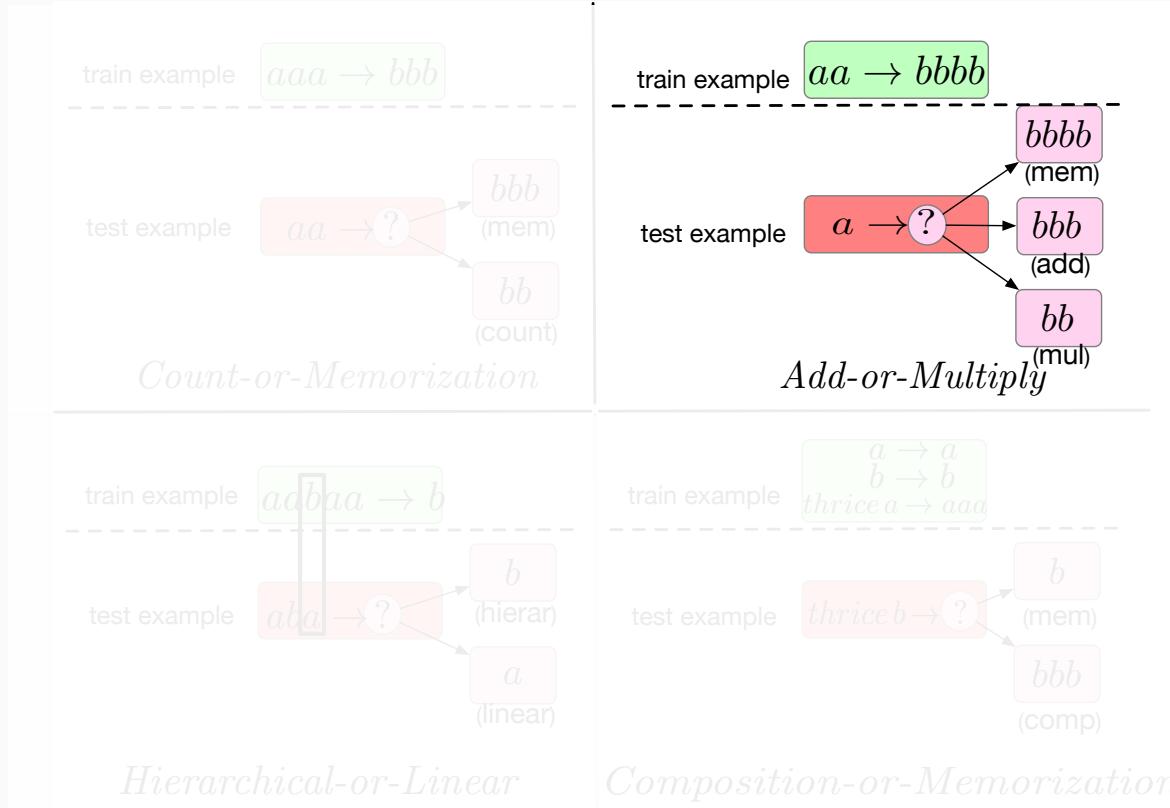
LSTMやTransformerがNLPで比較的うまく動くのは、階層的な帰納バイアスを持っているから？

※Transformerが階層的な処理を実現できるかは議論あり
[Hahn, 2020][Weiss+, 2021].

※この研究は経験的にどのような傾向を好むか(確率が比較的高く割り振られるか)を調べており、実現できるかは別問題

ニューラルネットの帰納バイアス

- 検証例2: Add, Multiply, or Memorize



小さい: その規則を好む

	length l	FPA ↑			L , nats ↓		
		add	mul	mem	add	mul	mem
LSTM-s2s no att.	20	0.00	0.94	0.00	25.42	0.31*	57.32
	15	0.07	0.65	0.00	19.24	4.67*	43.65
	10	0.95	0.01	0.00	0.68*	26.58	25.15
	5	0.04	0.00	0.00	17.12	50.83	18.60
LSTM-s2s att.	20	0.00	0.98	0.00	30.26	1.40*	58.84
	15	0.15	0.83	0.00	20.18	4.07*	46.36
	10	0.40	0.28	0.18	13.69	18.16	26.44
	5	0.00	0.00	0.97	45.88	77.86	0.01*
CNN-s2s	{5, 10, 15, 20}	0.00	0.00	1.0	> 318.12	> 346.19	0.00*
Transformer	{5, 10, 15, 20}	0.00	0.00	1.0	> 38.77	> 50.64	<3.50*

(b) Add-or-Multiply

CNNやTransformerはやはり暗記が得意

本研究がnature-nurture debateに与える影響 ニューラルネットは「空白の石版」ではない [Baroni, 2021]

- ニューラルモデル (生得的能力を持っていないままさらな状態と仮定) が、データのみからXを獲得するならば、Xは経験的に獲得できる 妥当?
 - "Our results also contribute to the long-running nature-nurture debate in language acquisition: whether the success of neural models implies that unbiased learners can learn natural languages with enough data, or whether human abilities to acquire language given sparse stimulus implies a strong innate human learning bias" [Papadimitriou & Jurafsky, 2020]

言語研究において

- deep netsは「どんな帰納バイアス (を持つモデル) が効率的な言語獲得に紐づくか」について分析するためのツールになれるといいのでは [Baroni, 2021]
 - この観点から、deep netsは言語獲得に関する algorithmic theoryとして捉えられても良いのでは [Baroni, 2021]
 - モデルの機能バイアスを(もれなく) probingすることも難しいと思うけど、人を調べる難易度よりはマシなのかな...

本当にそうなれるのかはおいておいて、こういう議論があるということは知っていても良いのでは

まとめ

- ニューラルモデルは特にアーキテクチャの観点で、汎化の選好(帰納バイアス)が大きく異なる
- 「ある現象Xをニューラルモデルがデータから学習できたので、Xは経験的に獲得できる知識である」という議論は短絡的
 - ニューラルネットと人に帰納バイアスの重複がまったくないことを仮定することになる
- 言語理論の立場から、ニューラル言語処理モデルを「どのような帰納バイアスが効率的な言語獲得と紐づくかを調べるツール」と考えれば良いのでは [Baroni, 2021]
(ということが議論されていることは頭の片隅に置いておいてよいのでは)

補足

テクニカルな面: 記述長に基づいた評価

方法論の特徴は大きく2つ

- どのような一般化を経験的に「好むか」を調べている

- [注意] あるモデルクラスが特定のパターンを認識する能力を持つかどうか (例えば, LSTMは $a^n b^n$ を受理できるか, TransformerはDyck- n 言語をモデリングできるか) のような話は別問題

- 学習の進む速さの軸を評価に取り入れている

$$L_{\mathcal{M}}(D) = - \sum_{t=2}^k \log p_{\mathcal{M}_{t-1}}(y_t | x_t) + c.$$

validation setを固定して, iterationごとのvalidation CEとかではだめだったか?

上の2つが要件であれば, 記述長 (と prequential codes) の話を持ち出す必然性はないと思う.

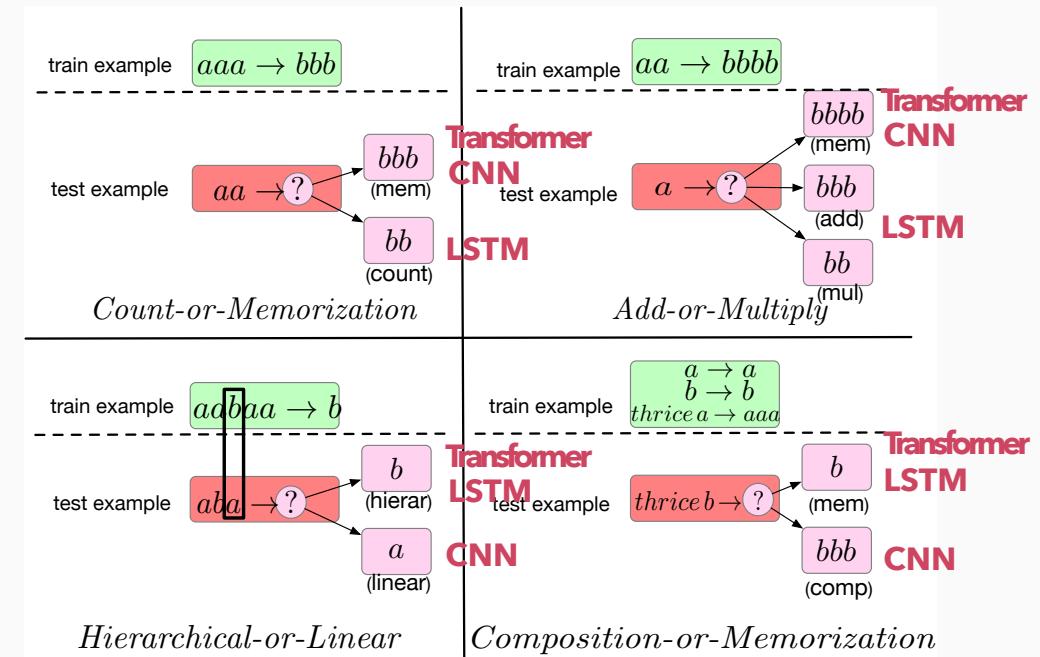
情報理論との接続はかっこいいのだが, そちらに厳密に合わせるせいで

不必要に複雑な実験手続きになっている印象.

各 t について, バッチを作り直してモデルを1から訓練し直す
規則ごとのtarget文字列の長さの違いは考慮しなくていいの?とか

ざっくり要約

- ニューラルseq2seqモデルは、アーキテクチャに応じて单一事例から異なる汎化を好む傾向
- つまり学習を始める前の段階でモデルアーキテクチャに応じて(ある意味生得的に、異なる)知識が入っている
 - 例えば階層的な汎化を好むなど人らしい帰納バイアスを有している場合がある
 - ある意味当たり前の議論。アーキテクチャを決めていること自体が帰納バイアスを仮定していること。
また、モデルごとに異なる帰納バイアスをもつことはリーダーボード的な研究からも明らか。



余談：画像処理モデルの帰納バイアス

- Identity Crisis: Memorization and Generalization under Extreme Overparameterization [Zhang+, 2020]
 - 入力をそのまま出力させるという学習を单一事例 (w/SGD) で行った時に、恒等変換を学習するか、その出力を覚えて出すことを学習するか。FCN, CNNで調査 (式の上ではいずれのモデルも恒等変換を学習することが可能)。
 - Linear, Non-Linear FCNsは单一事例から恒等変換を学習しない傾向

學習事例

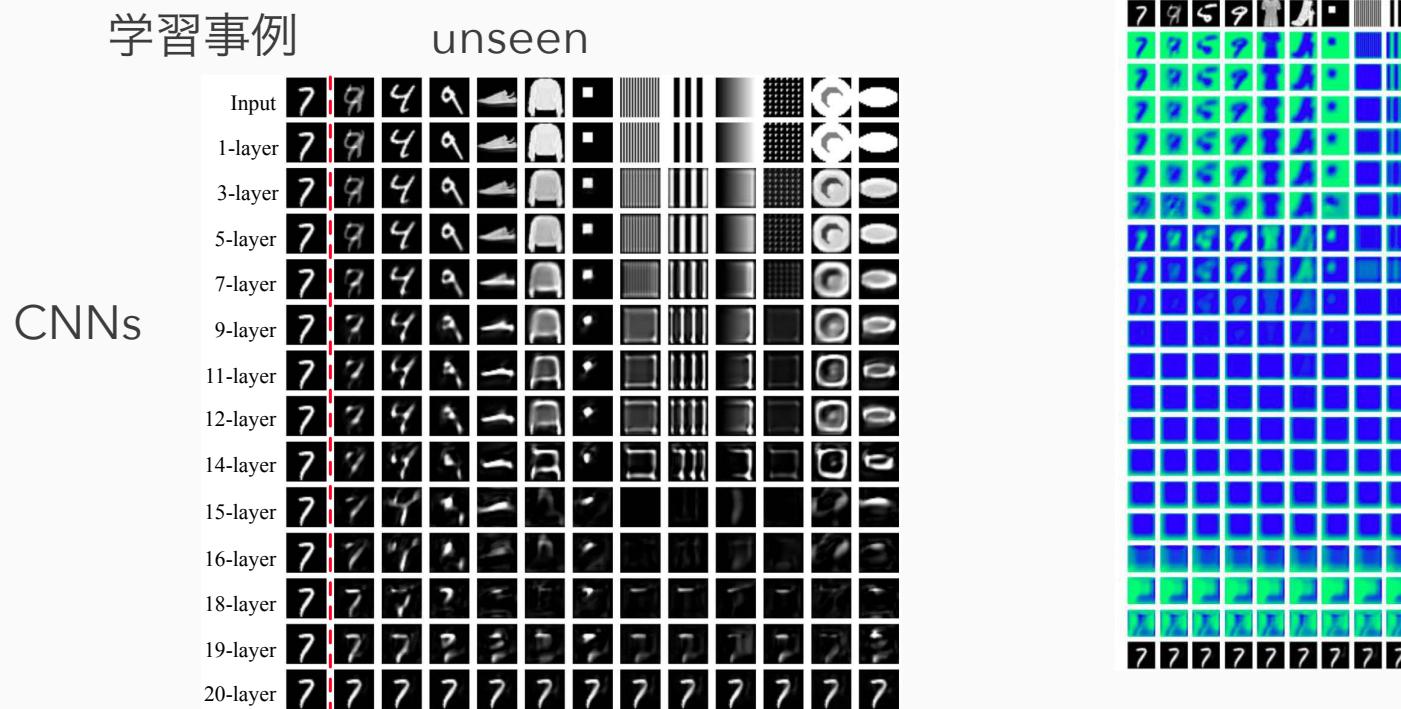
unseen

Linearを積んだだけなのに、
汎化傾向が変わる
(backpropの式が変わるため)

余談：画像処理モデルの帰納バイアス

- Identity Crisis: Memorization and Generalization under Extreme Overparameterization [Zhang+, 2020]

- 入力をそのまま出力させるという学習を一事例で行った時に、恒等変換を学習するか、その出力を覚えて出すことを学習するか。FCN, CNNで調査(式の上ではいずれのモデルも恒等変換を学習することが可能)。
- 浅いCNNsは单一事例から恒等変換を学習する傾向



深いネットワークは前半で入力をかき消し、後半層で特定の出力を再現させている可能性

(CNNは位置普遍な操作しかできないので、boundary (pad) 情報を駆使して出力を復元している可能性)

余談：ニューラルネットもshape-biasをもつ

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study [Ritter+, 2017]

Learning Inductive Biases with Simple Neural Networks [Feinman & Lake, 2018]

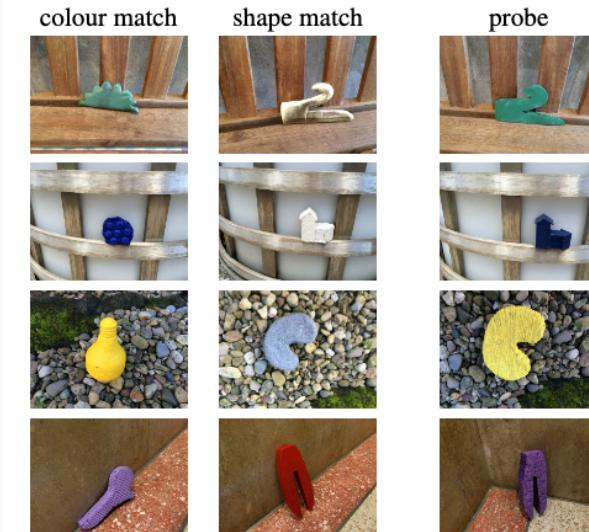
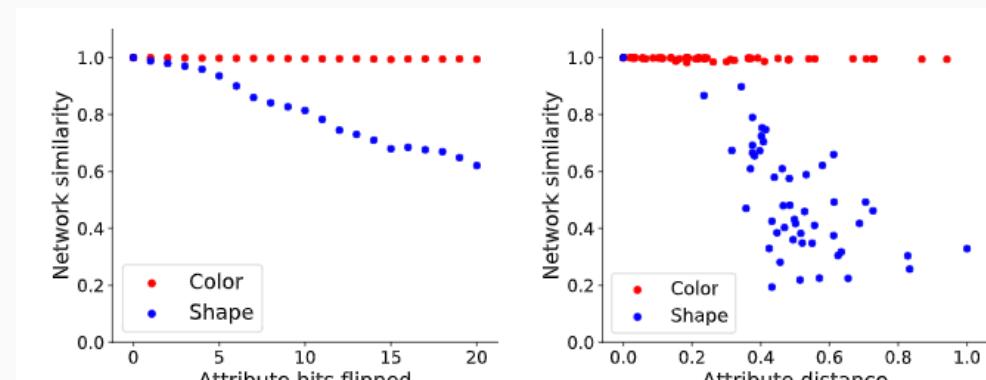
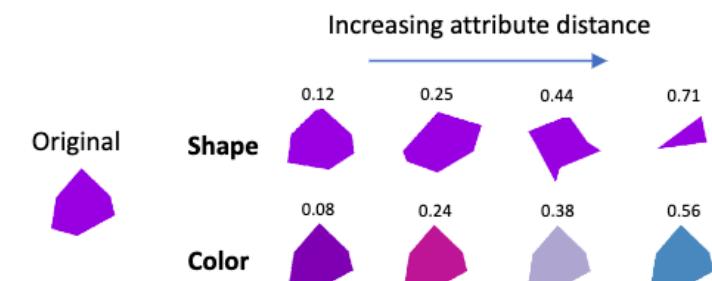


Figure 1. Example images from the Cognitive Psychology Dataset (see section 5). The data consists of image triples (rows), each containing a *colour match* image (left column), a *shape match* image (middle column) and a *probe* image (right column). We use these triples to calculate the shape bias by reporting the proportion of times that a model assigns the shape match image class to the probe image. This dataset was supplied by cognitive psychologist Linda Smith, and was designed to control for object size and background.



(a) MLP

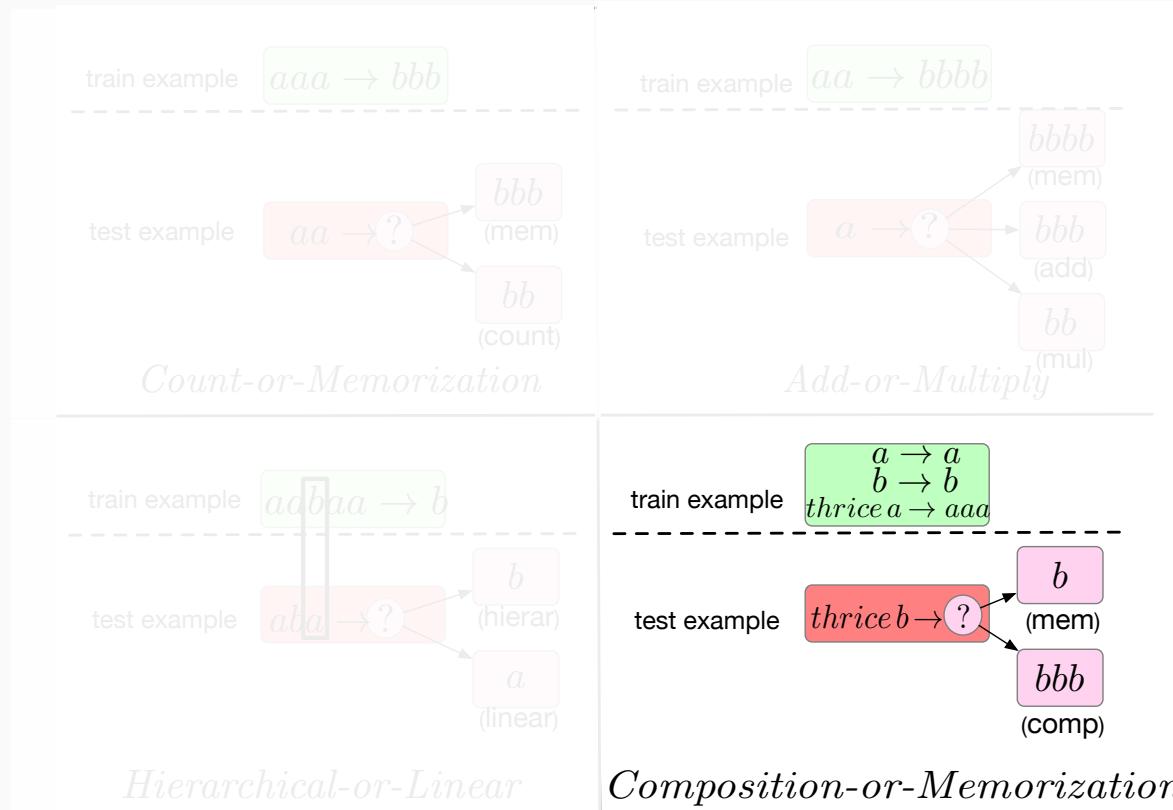
(b) CNN



(c) Distance along stimulus dimensions

ニューラルネットの帰納バイアス

- What they do when in doubt: a study of inductive biases in seq2seq learners [Kharitonov & Chaabouni, 2021]



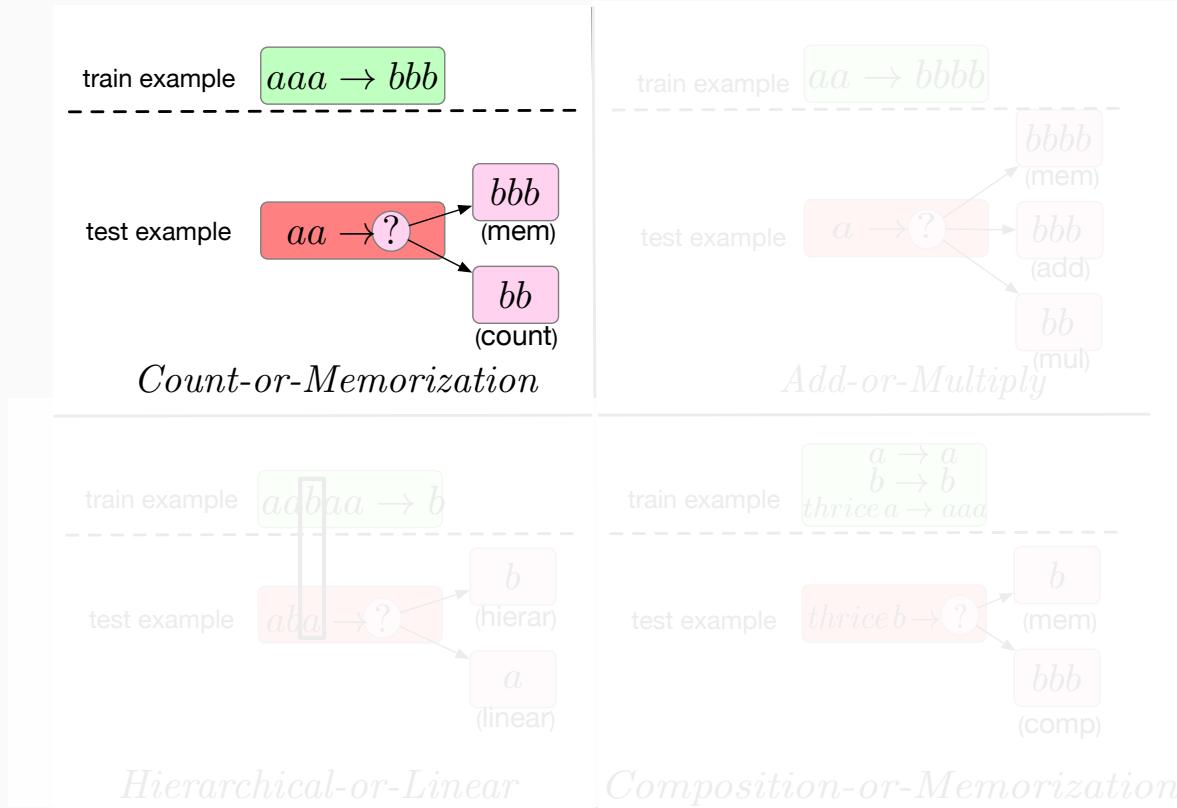
M , examples	FPA ↑		L , nats ↓	
	comp	mem	comp	mem
LSTM-s2s no att.	36	0.00	0.00	42.65 38.55
	24	0.00	0.00	238.54 89.36*
	6	0.00	0.00	656.93 157.55*
LSTM-s2s att.	36	0.00	0.00	62.34* 70.92
	24	0.00	0.00	263.33 157.82*
	6	0.00	0.00	659.85 164.43*
CNN-s2s	36	0.75	0.00	1.44* 49.92
	24	0.13	0.00	13.75* 84.55
	6	0.00	0.00	131.63 29.66*
Transformer	36	0.00	0.82	147.83 6.36*
	24	0.00	0.35	586.22 26.46*
	6	0.00	0.00	1235.01 53.91*

(d) Composition-or-Memorization

やっぱりTransformerが暗記する傾向
(Transformerがdata hungryであることに関連ある?)

ニューラルネットの帰納バイアス

- What they do when in doubt: a study of inductive biases in seq2seq learners [Kharitonov & Chaabouni, 2021]



	length l	FPA ↑		$L, \text{nats} ↓$	
		count	mem	count	mem
LSTM-s2s no att.	40	1.00	0.00	0.01*	97.51
	30	0.97	0.00	0.01*	72.67
	20	0.07	0.00	2.49*	55.67
	10	0.00	0.00	88.27	48.67*
LSTM-s2s att.	40	0.99	0.00	7.84*	121.48
	30	0.96	0.02	1.14*	83.48
	20	0.70	0.16	5.73*	49.33
	10	0.00	0.20	98.12	8.46*
CNN-s2s	{10, 20, 30, 40}	0.00	> 0.90	> 592.92	<1.31*
Transformer	{10, 20, 30, 40}	0.00	> 0.97	> 113.30	<11.14*

(a) Count-or-Memorization

CNNやTransformerは出力を暗記する傾向

(dropout率を上げるとmemの傾向は下がる。直感的。Appendix)

LSTMの数える能力は理論的な後付けあり [Weiss+, 2018]など