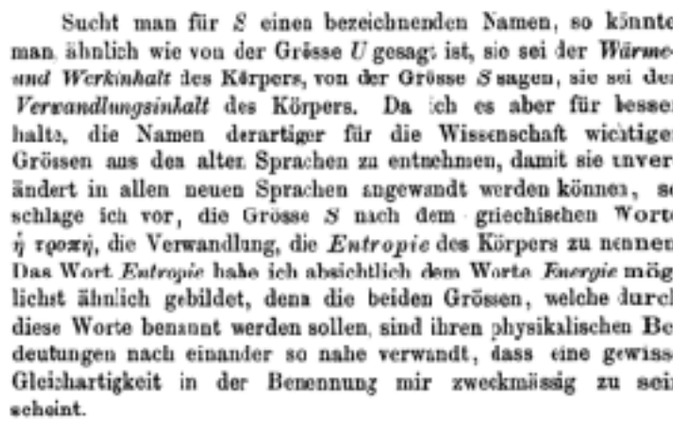
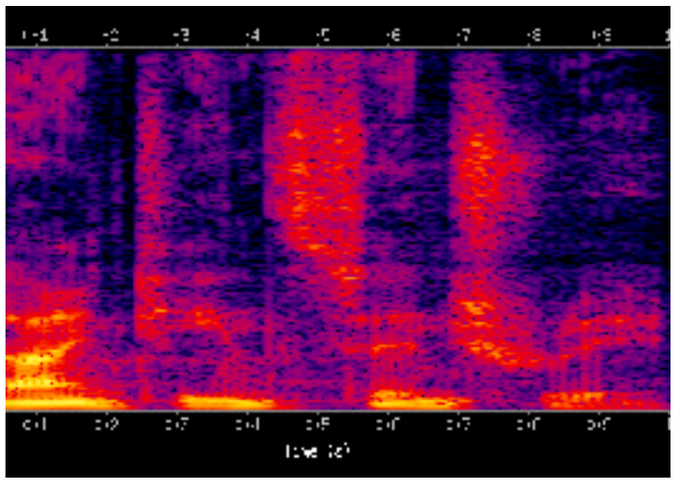
Multi-Task Learning:
One Model To Learn Them All

Aidan Gomez
L.Kaiser, N.Shazeer, A. Vaswani, N. Parmar, L. Jones and J. Uszkoreit

Domains of Research



Sucht man für S einen bezeichnenden Namen, so könnte man ähnlich wie von der Grösse U gesagt ist, sie sei der *Wärme- und Werkinhalt* des Körpers, von der Grösse S sagen, sie sei der *Verwandlungsinhalt* des Körpers. Da ich es aber für besser halte, die Namen derartiger für die Wissenschaft wichtiger Grössen aus den alten Sprachen zu entnehmen, damit sie unverändert in allen neuen Sprachen angewandt werden können, so schlage ich vor, die Grösse S nach dem griechischen Worte $\dot{\eta}$ τροπή, die Verwandlung, die *Entropie* des Körpers zu nennen. Das Wort *Entropie* habe ich absichtlich dem Worte *Energie* möglichst ähnlich gebildet, denn die beiden Grössen, welche durch diese Worte benannt werden sollen, sind ihren physikalischen Bedeutungen nach einander so nahe verwandt, dass eine gewisse Gleichartigkeit in der Benennung mir zweckmässig zu sein scheint.



Images: Google Images

Text

- RNNs
- Attention
- MoEs

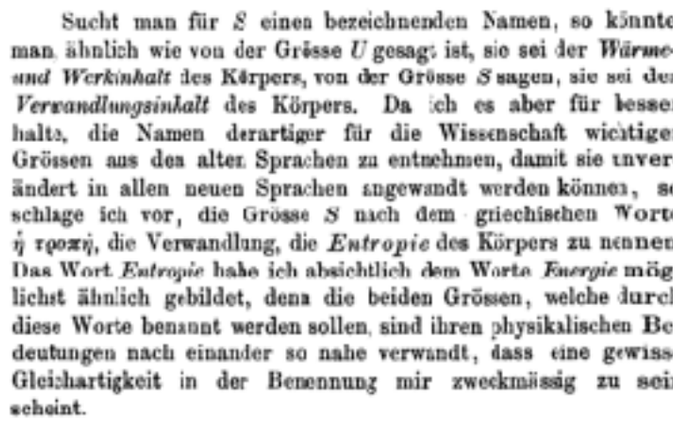
Audio

- RNNs
- Wavelet Transforms
- Hand-crafted features

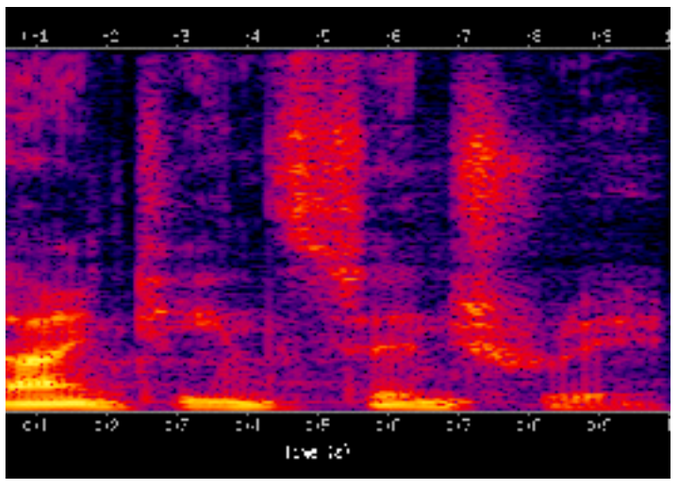
Vision

- Convolutions
- Residuals
- Separability

Domains of Research



Sucht man für S einen bezeichnenden Namen, so könnte man ähnlich wie von der Grösse U gesagt ist, sie sei der *Wärme- und Werkinhalt* des Körpers, von der Grösse S sagen, sie sei der *Verwandlungsinhalt* des Körpers. Da ich es aber für besser halte, die Namen derartiger für die Wissenschaft wichtiger Grössen aus den alten Sprachen zu entnehmen, damit sie unverändert in allen neuen Sprachen angewandt werden können, so schlage ich vor, die Grösse S nach dem griechischen Worte $\dot{\eta}$ τροπή, die Verwandlung, die *Entropie* des Körpers zu nennen. Das Wort *Entropie* habe ich absichtlich dem Worte *Energie* möglichst ähnlich gebildet, denn die beiden Grössen, welche durch diese Worte benannt werden sollen, sind ihren physikalischen Bedeutungen nach einander so nahe verwandt, dass eine gewisse Gleichartigkeit in der Benennung mir zweckmässig zu sein scheint.



Images: Google Images

Text

- RNNs
- Attention
- MoEs

Audio

- RNNs
- Wavelet Transforms
- Hand-crafted features

Vision

- Convolutions
- Residuals
- Separability

Each domain is tractable for:

<u>Auto-regressive ConvNets</u>

ByteNet
(Kalchbrenner et al. 2016)

WaveNet
(Van Der Oord et al. 2016)

PixelCNN
(Van Der Oord et al. 2016)

Auto-Regressive Convolutional Networks

PixelCNN: Image Generation

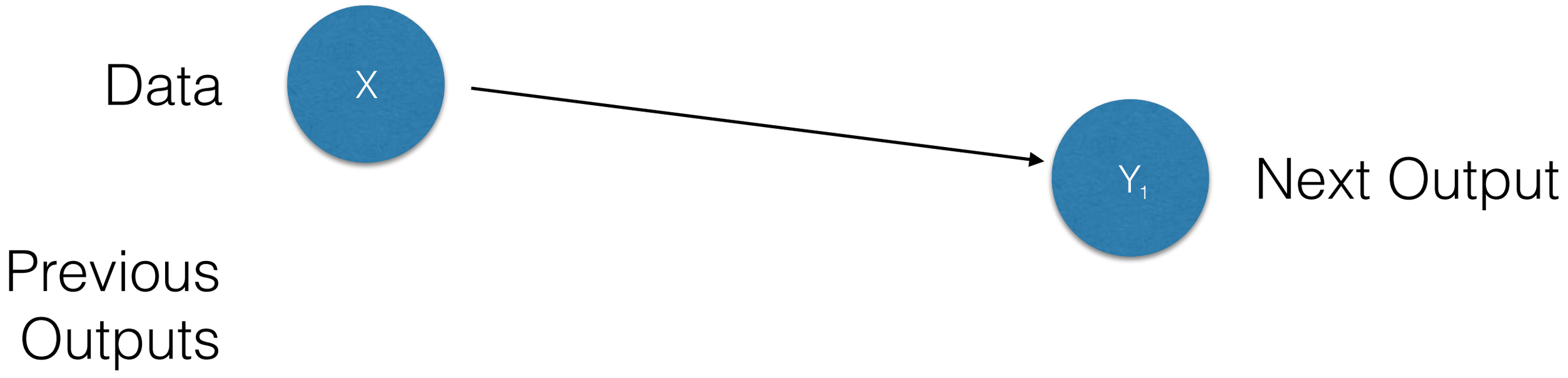
WaveNet: Text to Speech

ByteNet: Translation

Matt Casaday, 25, a senior at Brigham Young University, says he had paid 42 cents on Amazon.com
for a used copy of “Strategic Media Decisions: Understanding The Business End Of The Advertising Business.”

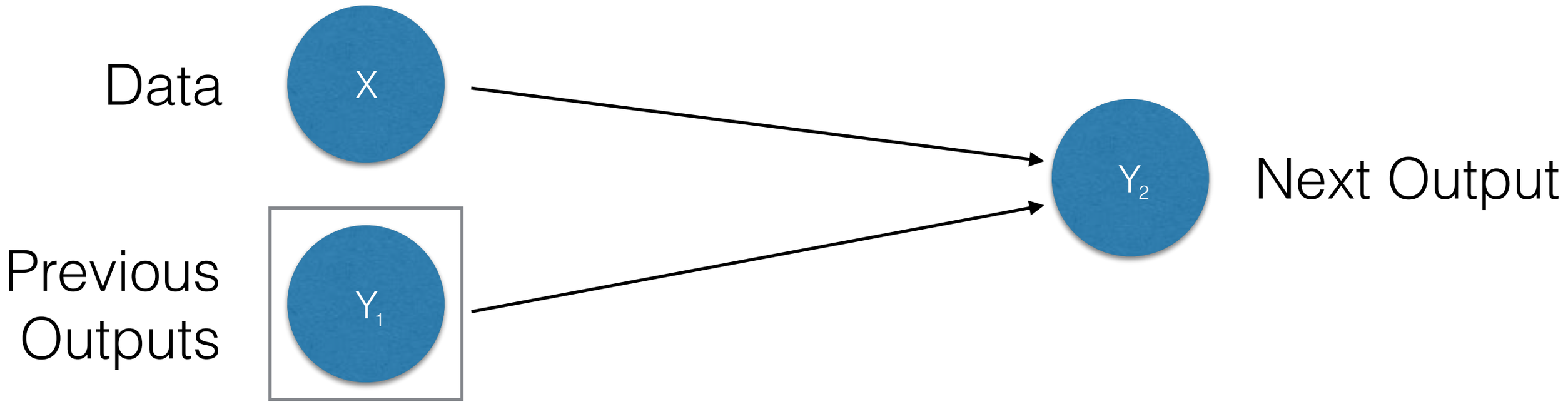
Matt Casaday, 25, ein Senior an der Brigham Young University, sagte, er habe 42 Cent auf Amazon.com
für eine gebrauchte Kopie von “Strategic Media Decisions: Understanding The Business End Of The Advertising Business”.

Auto-Regressive Models



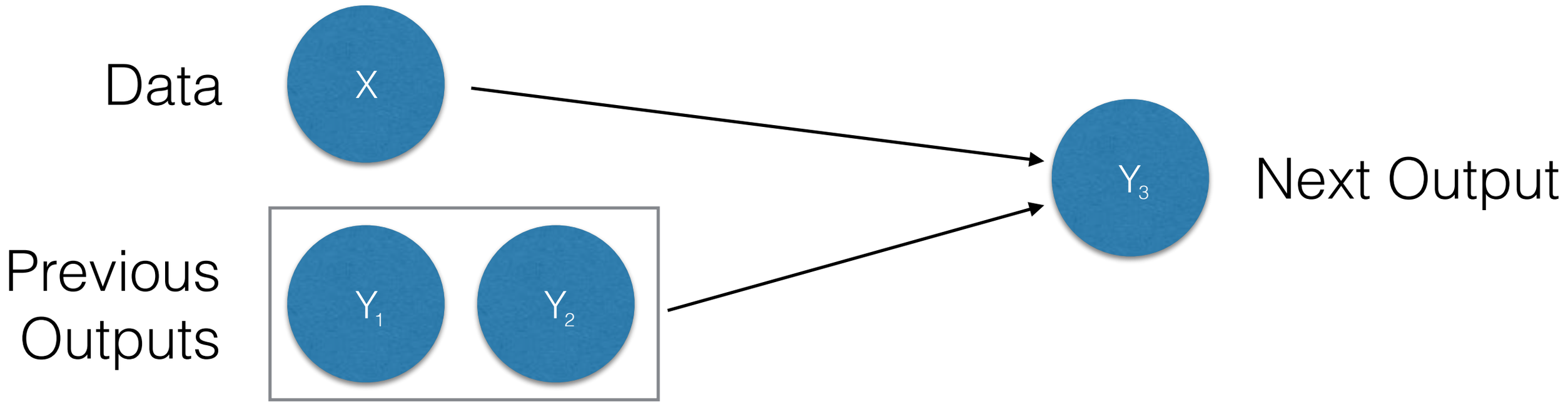
Future outputs conditioned on all past outputs.

Auto-Regressive Models



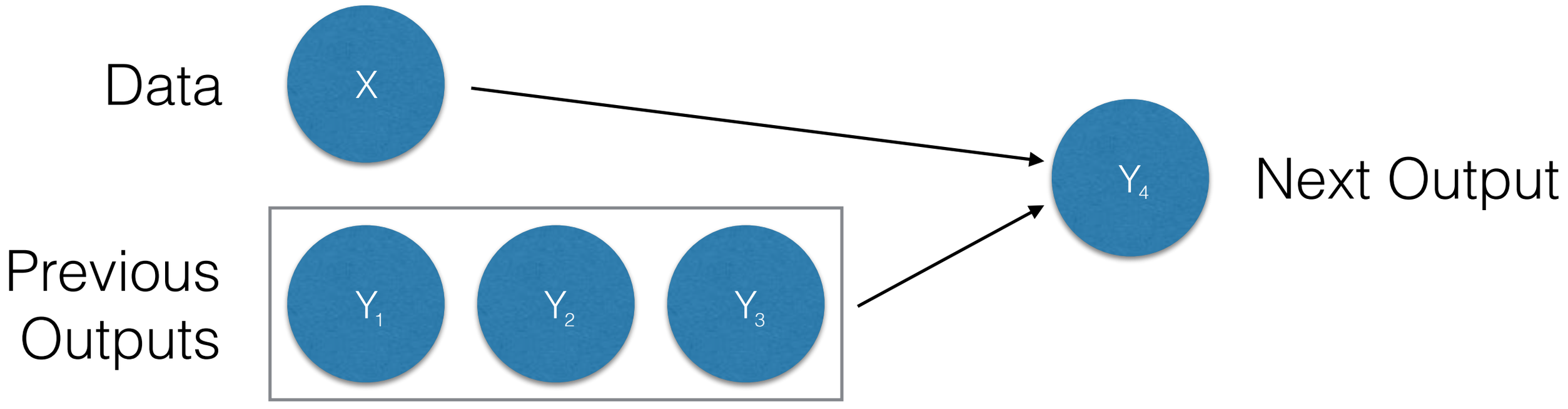
Future outputs conditioned on all past outputs.

Auto-Regressive Models



Future outputs conditioned on all past outputs.

Auto-Regressive Models



Future outputs conditioned on all past outputs.

Questions

- Will some ‘Grand Unified Model’ be able to match SOTA performance across many domains?

- Will “tricks” developed for particular domains boost or harm performance in others?

- Will a unified model better-demonstrate transfer learning and few-shot learning, analogous to humans?

To English

"A man that is sitting in front of a suitcase"

To Category

Category 127 (Male Human)

"Last week, Kigali raised the possibility of military retaliation after shells..."

To French

"La semaine dernière, Kigali a soulevé la possibilité de représailles militaires aprés avoir débarqué des coquilles..."

"Can you give our readers some details on this?"

To German

"Können Sie unseren Lesern einige Details dazu geben?"

The above represents a triumph of either apathy or civility

To Parse

"S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S"

The Goal

Take data from any domain and have the model perform any task and output the result to any desired domain.

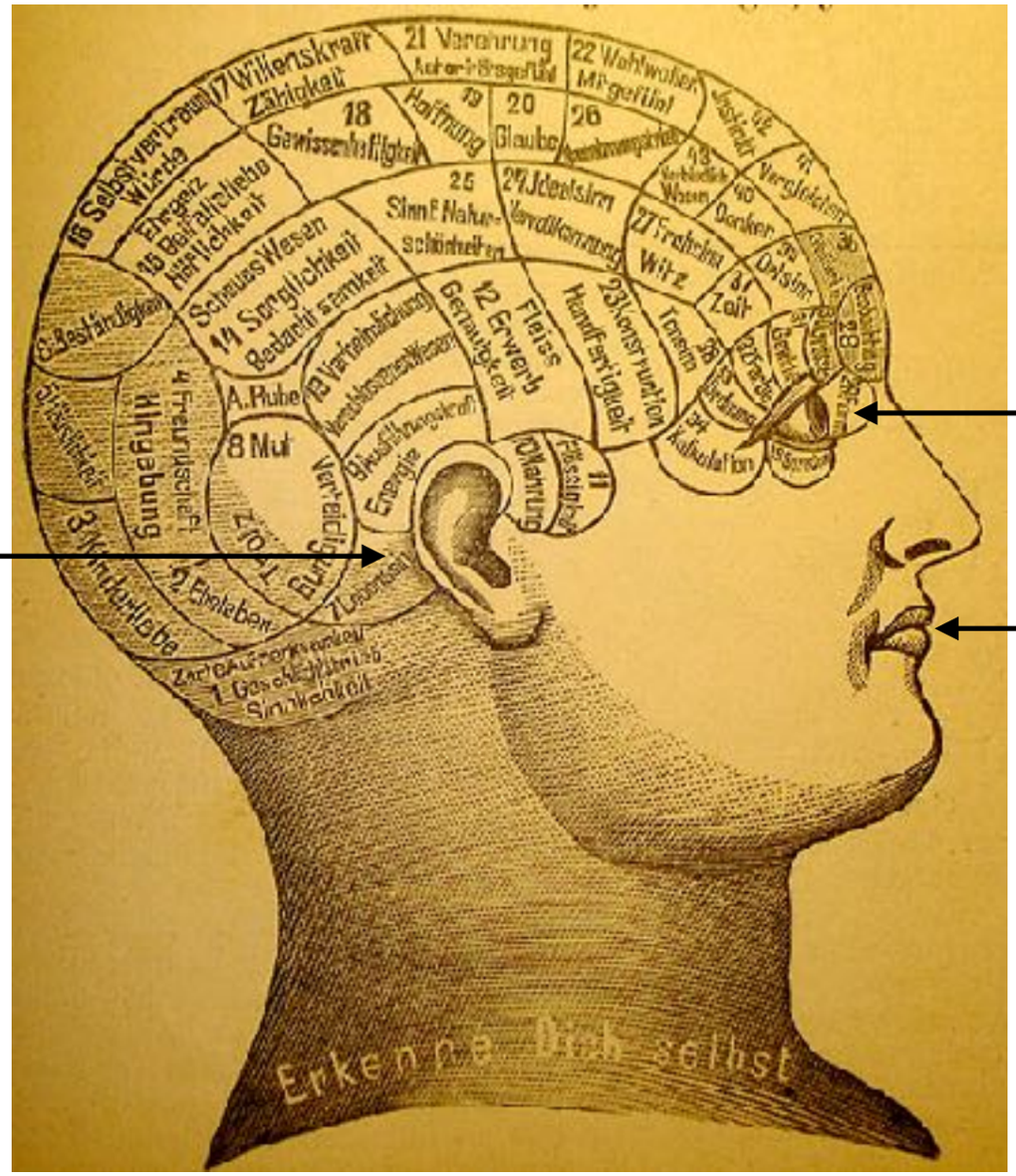
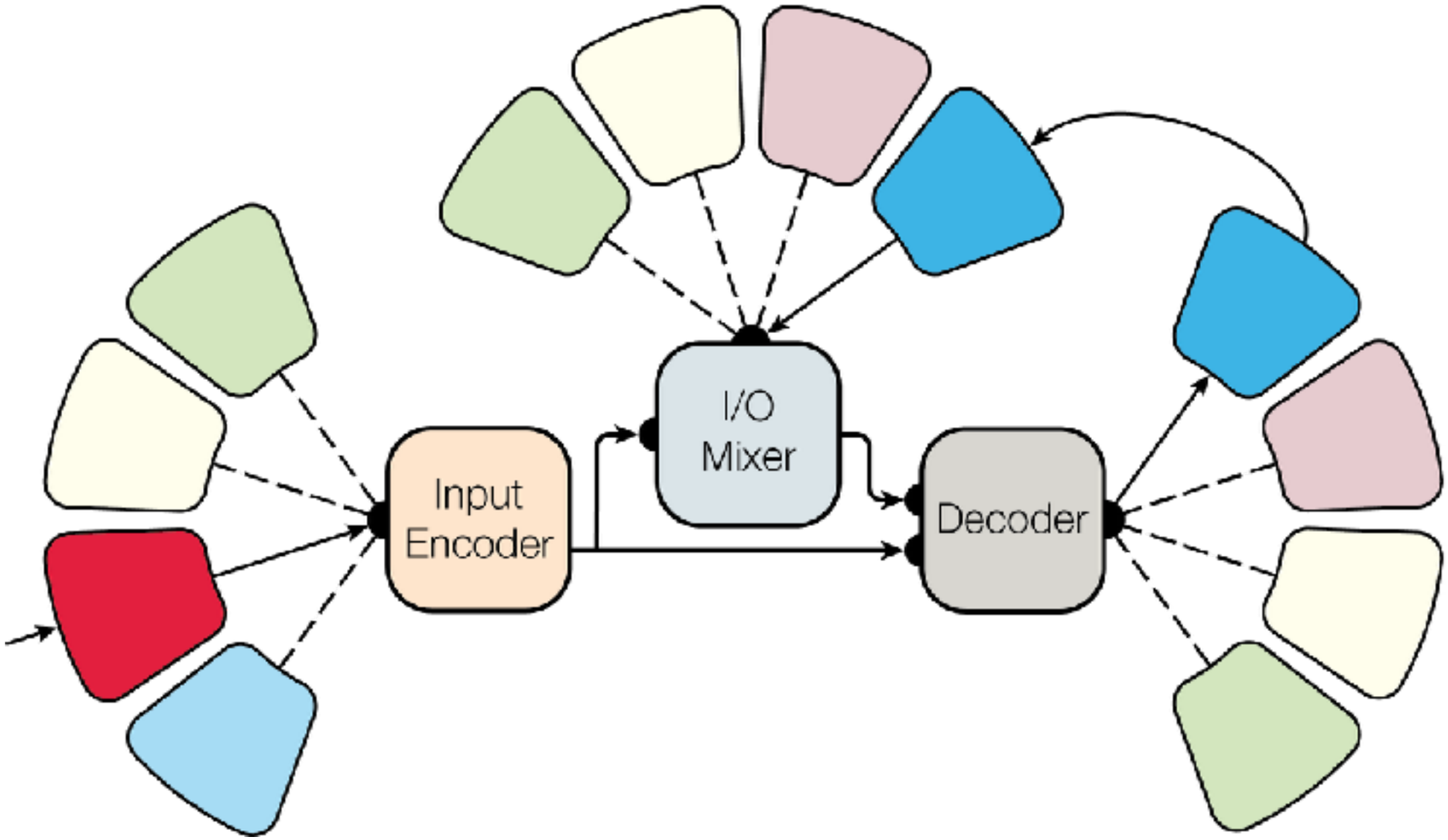


Image: Wikipedia

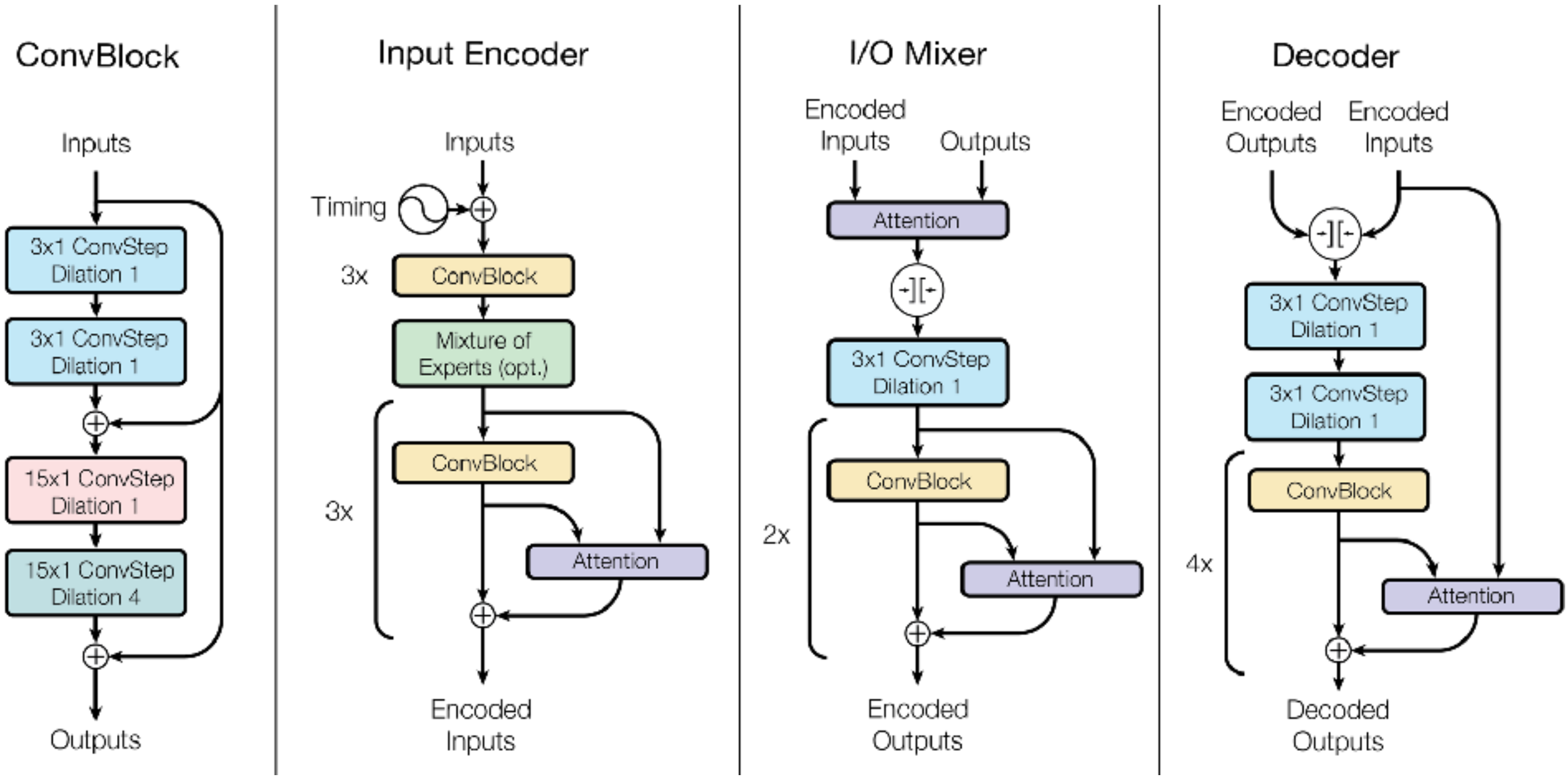
Modalities

Interfaces between the external world and the internal mind



Our Model

Convolutional Auto-Regressive Sequence Model



Our Model

Convolutional Auto-Regressive Sequence Model

Our Model

- Internally, our model consists of convolutional blocks, attention and sparsely-gated mixture of experts (MoE).

- Draws general architecture from Xception (Chollet, 2016), modified to support auto-regression, MoEs and attention.

- It learns to solve 8 problems simultaneously (on a single parameter set).

- Demonstrates strong transfer learning.

- Performance exceeds common baselines with no tuning whatsoever.

The Problems

- ImageNet Image Classification
- WMT English -> German Translation
- WMT German -> English Translation
- WMT English -> French Translation
- WMT French -> English Translation
- MSCOCO Image Captioning
- WSJ1 English Audio Transcription
- Penn Treebank Grammar Parsing

Transfer Learning

Problem	Joint (8 Problems)	Alone
ImageNet	76%	77%
WSJ (Speech -> Text)	41%	23%
WMT (Eng -> Ger)	72%	71%
Grammar Parsing	14.5%	11.7%

Performance on tasks with limited data (WSJ, Parsing) were substantially improved when trained alongside 7 other problems.

Multi-Task as a Regularizer

Problem	With ImageNet	Alone
Grammar Parsing	12.7%	11.7%

Two tasks that are seemingly without relation confirms:

- Any prior is better than no prior
- The best regularizer is more data (even if it has nothing to do with the task at hand, apparently)

Crucial Components

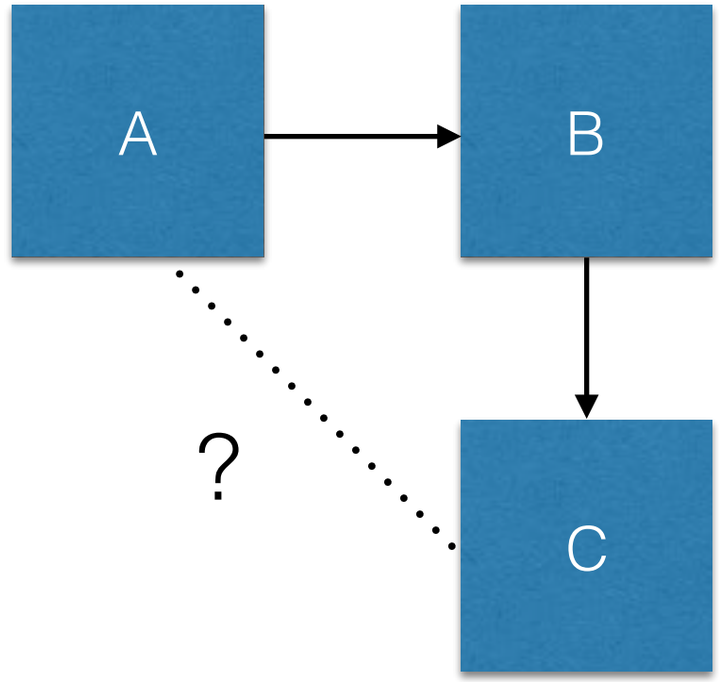
Problem	All Blocks	No Attention	No MoE
ImageNet	67%	67%	66%
WMT (EN-FR)	76%	72%	74%

- Attention clearly improves performance on NMT tasks and doesn't seem to harm performance on vision tasks.

- Mixture of Experts (MoE) may be valuable in models for computer vision (previously only applied to NMT).

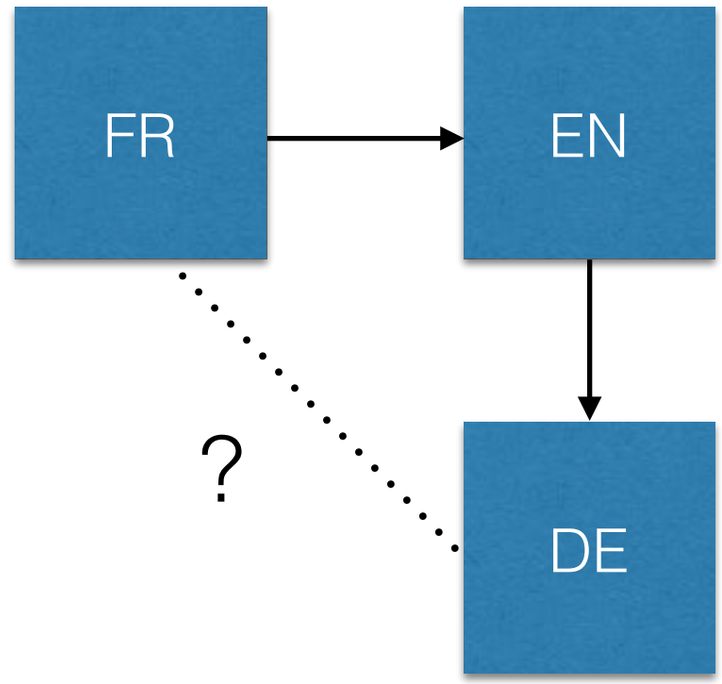
Zero-Shot Learning

Can our model learn to simplicity infer tasks that it has never received training data for?



Zero-Shot Learning

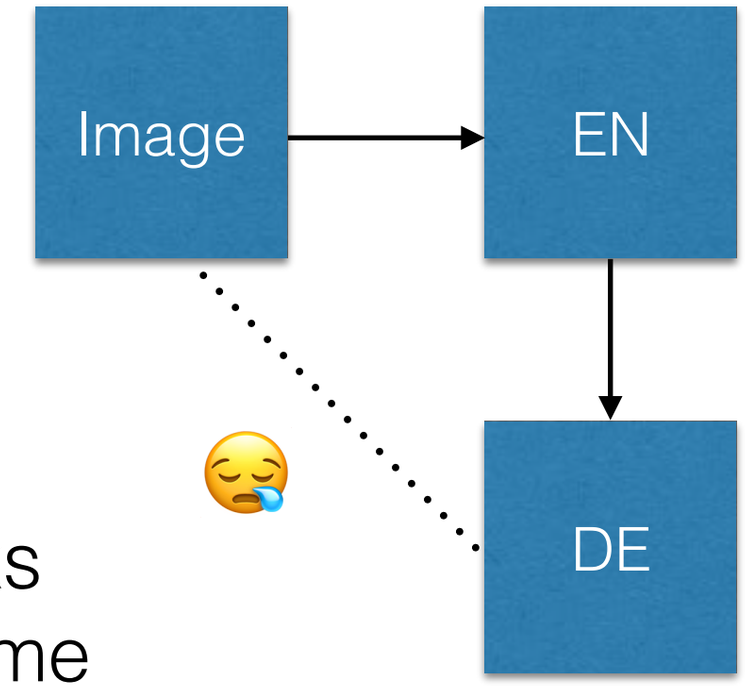
Can our model learn to simplicity infer tasks that it has never received training data for?



Yes! But…
It has been shown that models seem to naturally demonstrate this property intra-modality (Wu et al., 2016).
Unfortunately, inter-modality appears to be a more difficult problem…

Domain Adaptation

- The problems associated with inter-domain zero-shot learning stem from poor domain adaptation.



- Distributions of data for different tasks are dissimilar (even when from the same domain), enabling the model to infer the task it expects to perform on the data.

- Techniques with potential to improve domain adaptation in our model: domain adversaries, representation similarity rewards, others?

Future Work

- Generative Tasks?
- Cross-domain zero-shot learning?
- Performance closer to SOTA?
- Help build T2T!

T2T: Tensor2Tensor

- Released today!
- GitHub: tensorflow/tensor2tensor
- New open-source project released by Google Brain built by the authors of this project.
- For auto-regressive Tensor2Tensor models.
- WaveNet, Neural GPU, ByteNet, FAST, SliceNet, MultiModal (this work) come built in!

Related Work

- Attention Is All You Need: https://arxiv.org/abs/1706.03762

- Depthwise Separable Convolutions for Neural Machine Translation: https://arxiv.org/abs/1706.03059