




One Model To Learn Them All

Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish
Vaswani, Niki Parmar, Llion Jones, Jakob Uszkoreit



CS546 Course Presentation
Shruti Bhargava (shrutib2)
Advised by : Prof. Julia Hockenmaier

Outline

- **Motivation**
- Understanding the task
- Model Architecture
- Datasets
- Training details
- Performance Evaluation
- Key contributions/ Limitations

Motivation

1. Process the question and think of an answer
2. Convey the answer to me

What is your favourite fruit?

Write?

Draw?

Speak?

Apple



/'apəl/

Text
Modality

Image
Modality

Audio
Modality

Motivation

- Humans reason about concepts independent of input/output modality
- Humans are able to reuse conceptual knowledge in different tasks

Understanding the task

- **Multimodal Learning:** single task, different domains

Eg. Visual Question Answering

Input: Images + Text, Output: Text

- **Multitask Learning:** multiple tasks, mostly same domain

Eg. Translation + Parsing

- This work = **Multimodal + Multitask**



Question addressed :

Can one unified model solve tasks across
multiple domains?



Multiple Tasks/Domains, One Model - MultiModel



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

“Can you give our
readers some details
on this?”

The above represents
a triumph of either
apathy or civility

To English

To Category

To French

To German

To Parse

“A man that is
sitting in front of
a suitcase”

Category 127
(Male Human)

“La semaine dernière,
Kigali a soulevé la
possibilité de
représailles militaires
après avoir débarqué
des coquilles...”

“Können Sie unseren
Lesern einige
Details dazu geben?”

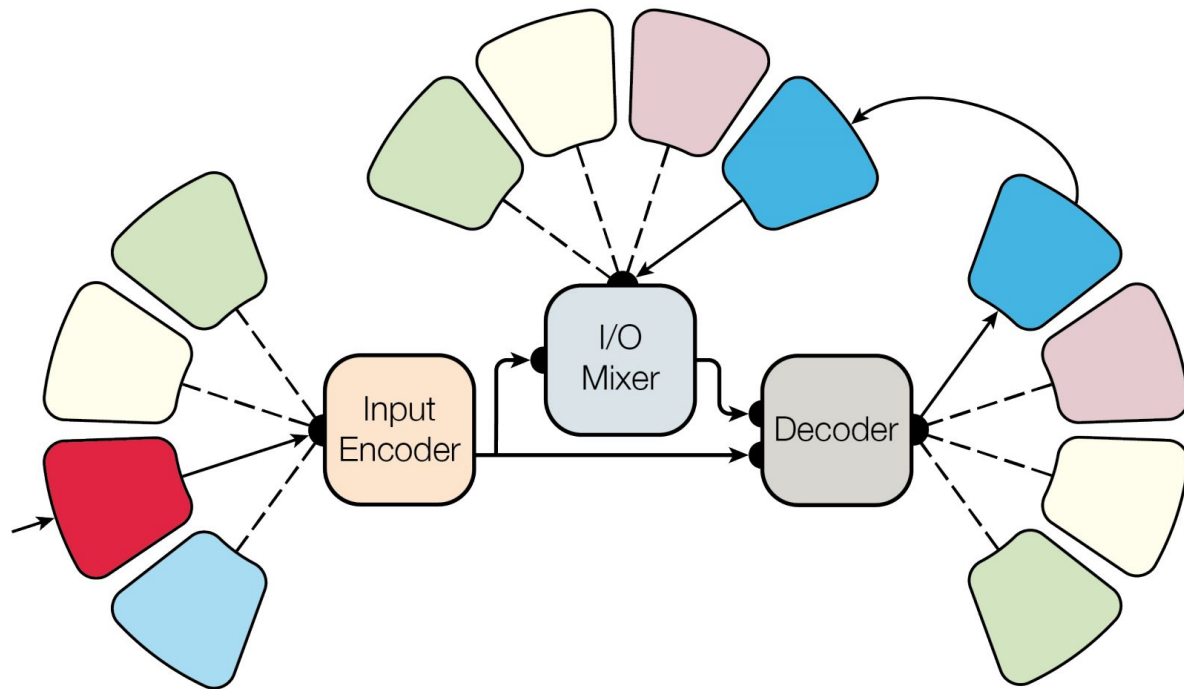
“S NP DT JJS /NP
VP VBZ NP NP DT
NN /NP PP IN NP
NP NN /NP CC NP
NN /NP /NP /PP /NP
/VP . /S”

Outline

- Motivation
- Understanding the task
- **Model Architecture**
- Datasets
- Training details
- Performance Evaluation
- Key contributions / Limitations

MultiModel Architecture

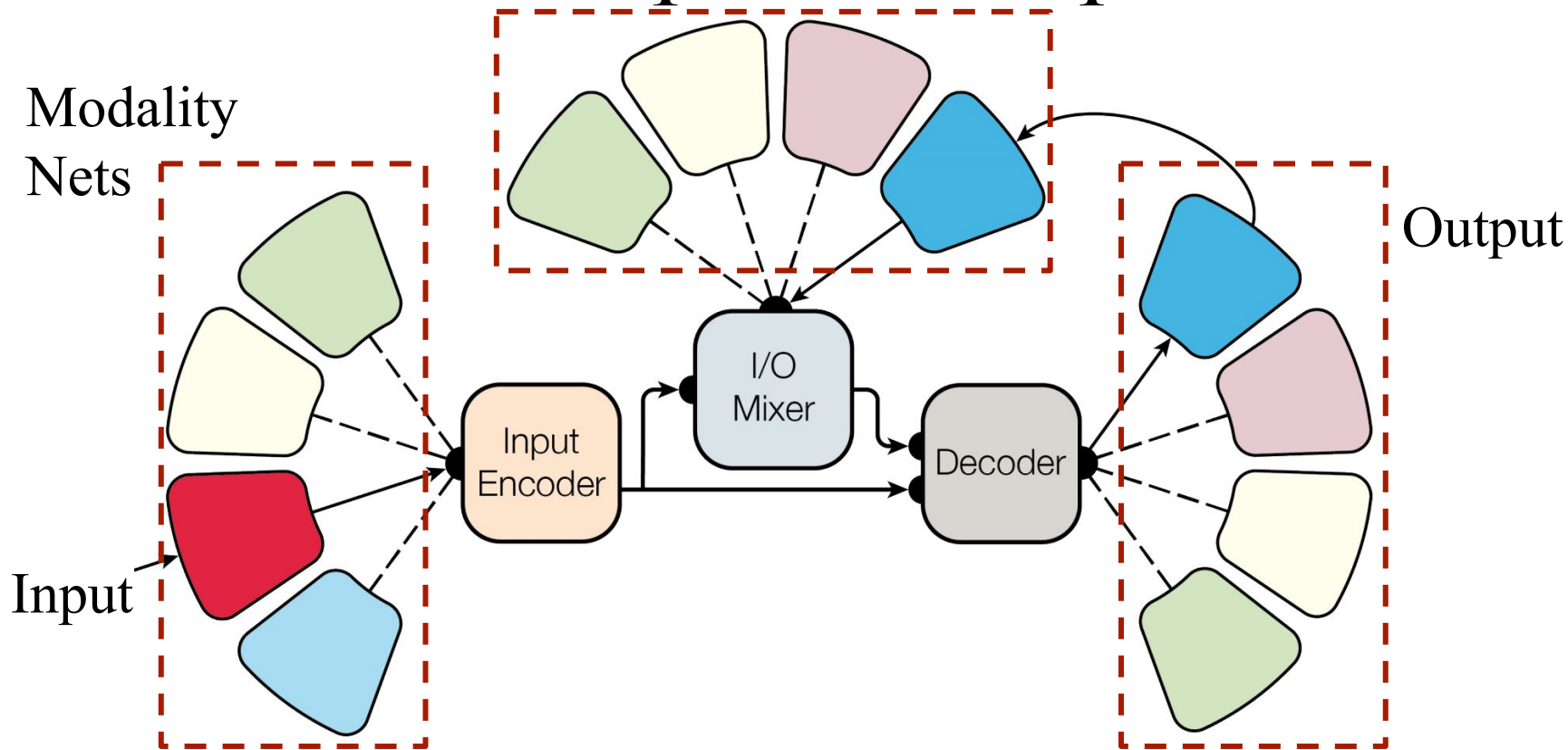
- Modality Nets
- Encoder-Decoder
- I/O Mixer



MultiModel: Input \rightarrow Output

- **Modality Net:** domain-specific input \rightarrow unified representation
- **Encoder:** unified input representations \rightarrow encoded input
- **I/O Mixer:** encoded input \Rightarrow previous outputs
- **Decoder:** decodes (input + mixture) \rightarrow output representation
- **Modality Net:** unified representation \rightarrow domain-specific output

MultiModel: Input \rightarrow Output

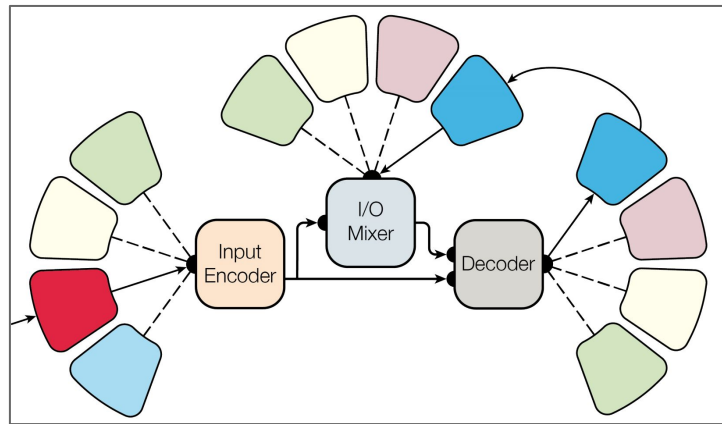


MultiModel: Modality Nets

Domain-specific Representation \leftrightarrow Unified Representation

4 modality nets - One net per domain

- Language
- Image
- Audio
- Categorical - only output



Modality Nets: Language Modality

Input tokenized using 8k subword units

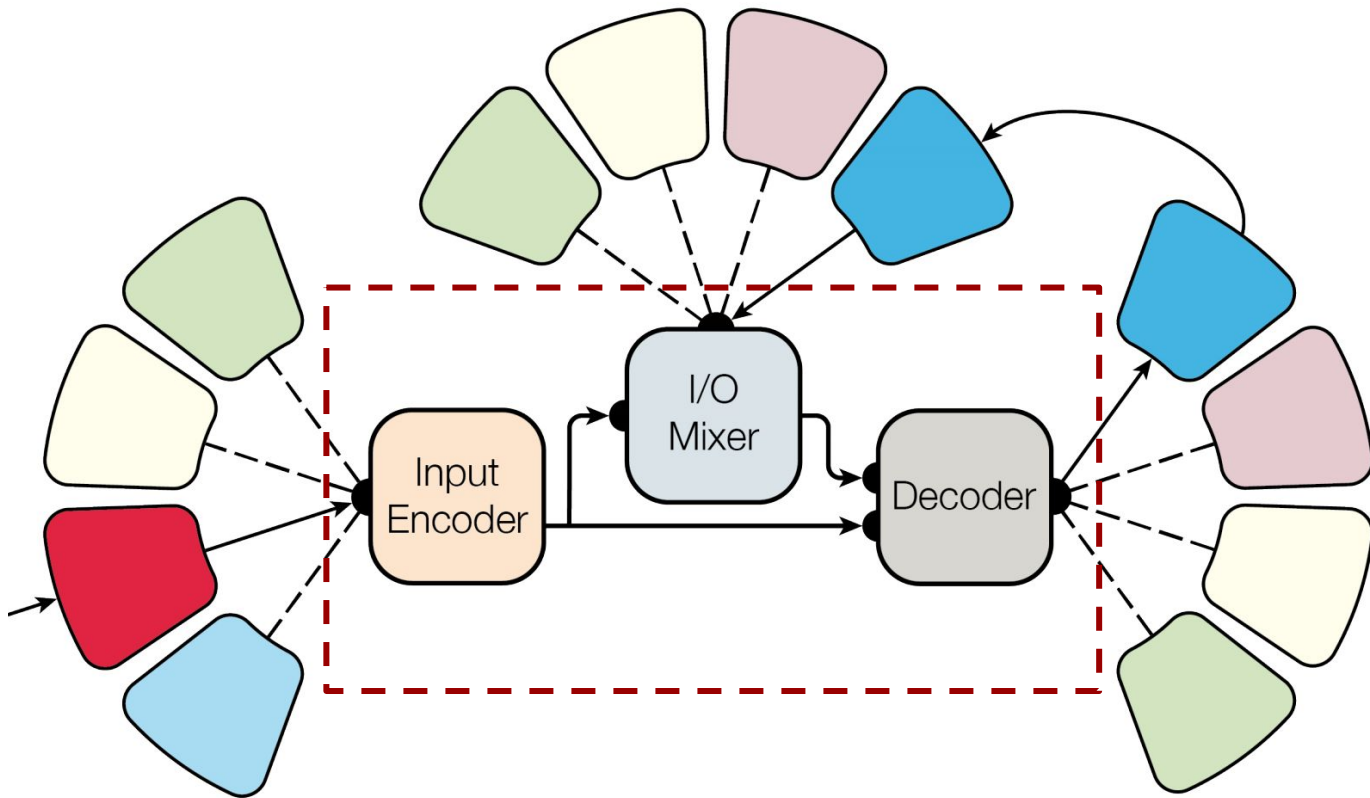
- Acts as an open vocabulary example - [ad|mi|ral]
- Accounts for rare words

Input Net - $LanguageModality_{in}(x, W_E) = W_E \cdot x$

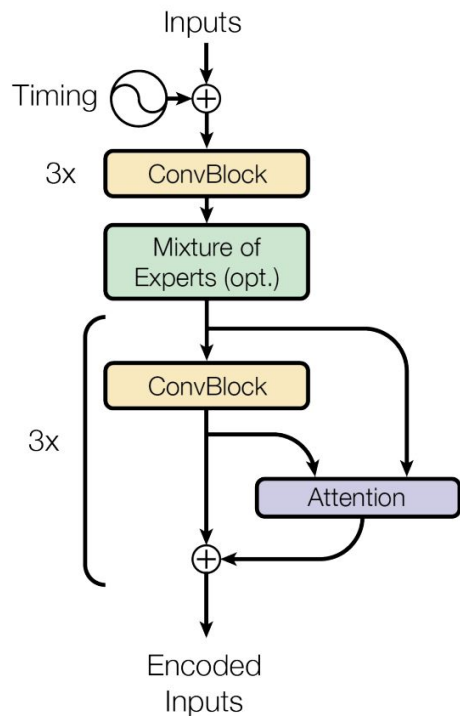
Output Net - $LanguageModality_{out}(x, W_S) = Softmax(W_S \cdot x)$

See Details for Vocabulary construction [here](#).

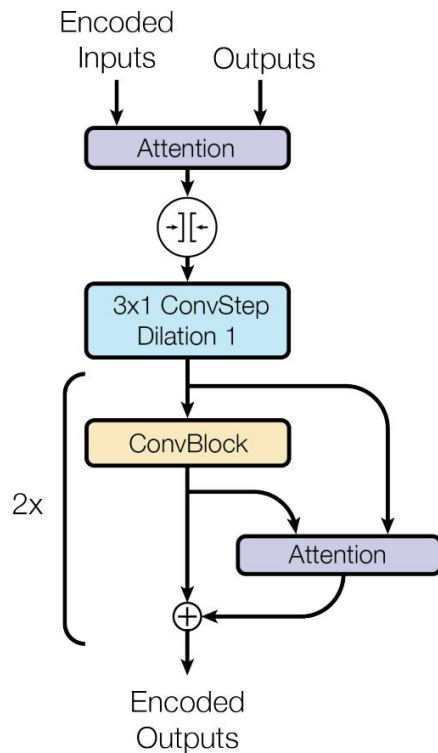
MultiModel: Domain Agnostic Body



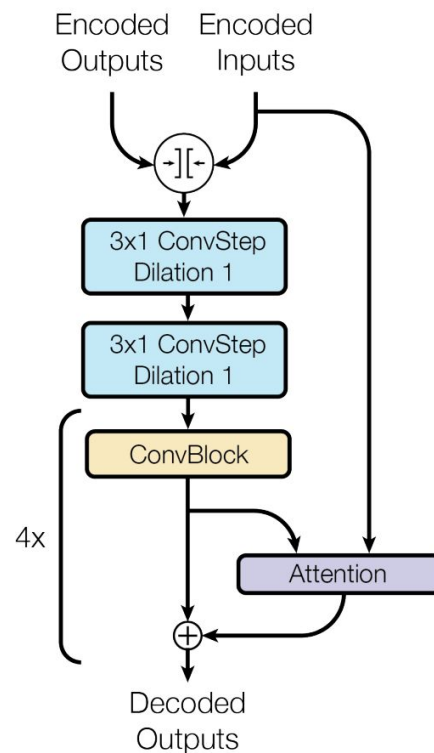
MultiModel: Domain Agnostic Body



Input Encoder



I/O Mixer



Decoder

MultiModel: Building Blocks

Combines 3 state-of-the-art blocks:

- Convolutional: SOTA for images
- Attention: SOTA in language understanding
- Mixture-of-Experts (MoE): studied only for language

Building Block: ConvBlock

$$\text{ConvStep}_{d,s,f}(W, x) = \text{LN}(\text{SepConv}_{d,s,f}(W, \text{ReLU}(x))).$$

Depthwise Separable Convolutions

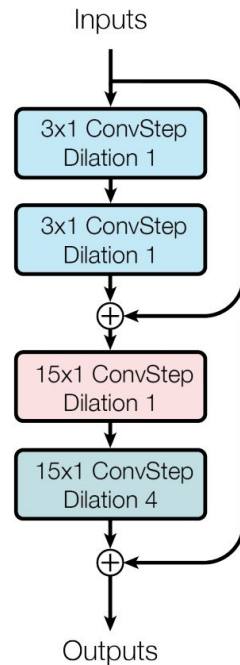
- convolution on each feature channel
- pointwise convolution for desired depth.

Layer Normalisation

- Statistics computed for a layer (per sample)

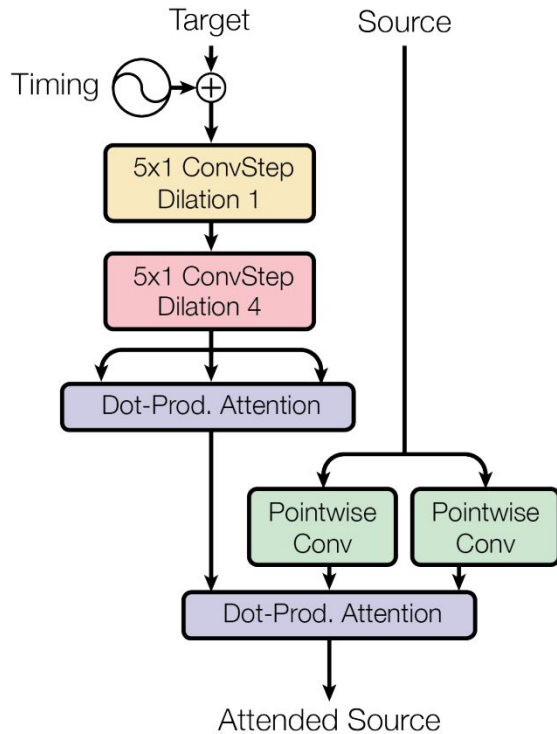
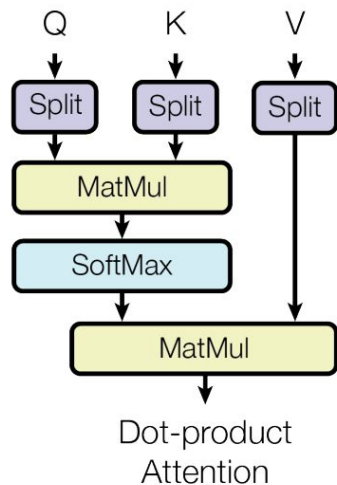
See Details on [Layer normalisation](#) and [Separable Convolutions](#).

ConvBlock



Building Block: Attention

Dot-Prod. Attention



See Details on the attention block [here](#).

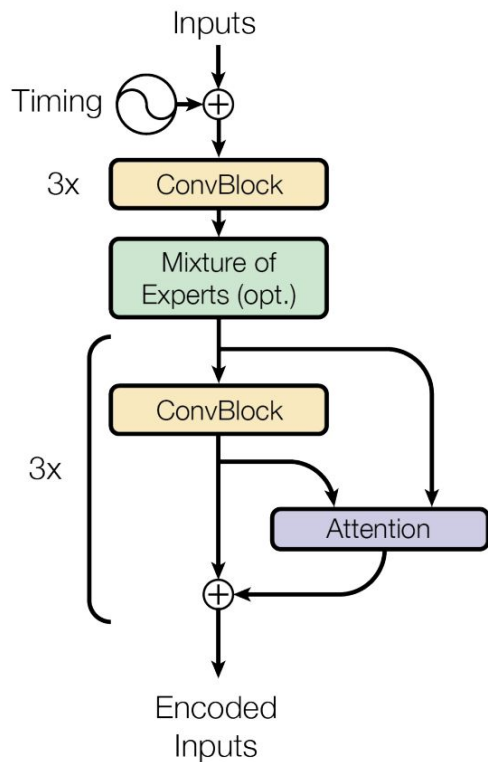
Building Block: Mixture of Experts

Sparsely-gated mixture-of-experts layer

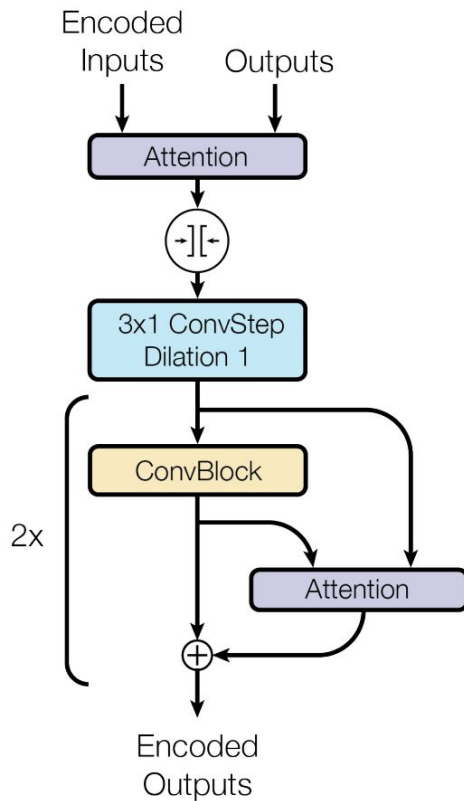
- Experts: feed-forward neural networks
- Selection: trainable gating network
- Known booster for language tasks

See Details on the MoE block [here](#).

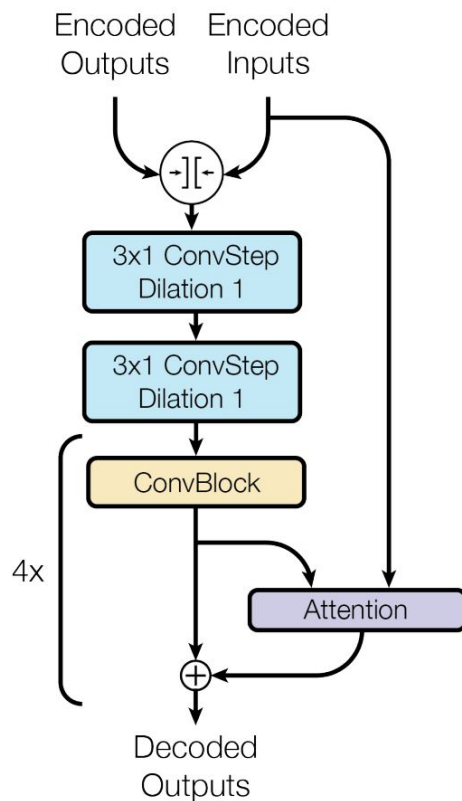
Input Encoder



I/O Mixer



Decoder



Structurally similar to Bytenet, read [here](#)

Outline

- Motivation
- Understanding the task
- Model Architecture
- **Datasets**
- Training details
- Performance Evaluation
- Key contributions / Limitations

Datasets/Tasks

- WSJ speech
- WSJ parsing
- ImageNet
- COCO image-captioning
- WMT English-German
- WMT German-English
- WMT English-French
- WMT German-French

Outline

- Motivation
- Understanding the task
- Model Architecture
- Datasets
- **Training details**
- Performance Evaluation
- Key contributions / Limitations

Training Details

- Token for task eg. *To-English* or *To-Parse-Tree*, to decoder.
Embedding vector for each token learned.
- Mixture of experts block :
 - 240 experts for joint training, 60 for single training
 - Gating selects 4
- Adam optimizer with gradient clipping
- Experiments on all tasks use same hyperparameter values

Outline

- Motivation
- Understanding the task
- Model Architecture
- Datasets Used
- Training details
- **Experiments/ Results**
- Key contributions / Limitations

Experiments

- MultiModel vs state-of-the-art?
- Does simultaneous training on 8 problems help?
- Blocks specialising in one domain help/harm other?

Results

1. MultiModel vs state-of-the-art ?

Problem	MultiModel (joint 8-problem)	State of the art
ImageNet (top-5 accuracy)	86%	95%
WMT EN \rightarrow DE (BLEU)	21.2	26.0
WMT EN \rightarrow FR (BLEU)	30.5	40.5

Results

2. Does simultaneous training help?

Problem	Joint 8-problem		Single problem	
	log(perplexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.7	66%	1.6	67%
WMT EN→DE	1.4	72%	1.4	71%
WSJ speech	4.4	41%	5.7	23%
Parsing	0.15	98%	0.2	97%

Problem	Alone			W/ ImageNet			W/ 8 Problems		
	log(ppl)	acc.	full	log(ppl)	acc.	full	log(ppl)	acc.	full
Parsing	0.20	97.1%	11.7%	0.16	97.5%	12.7%	0.15	97.9%	14.5%

Results

3. Blocks specialising in one domain help/harm other?

MoE, Attention - language experts

Problem	All Blocks		Without MoE		Without Attention	
	log(perplexity)	accuracy	log(perplexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.6	67%	1.6	66%	1.6	67%
WMT EN→FR	1.2	76%	1.3	74%	1.4	72%

Outline

- Motivation
- Understanding the task
- Model Architecture
- Datasets Used
- Training details
- Performance Evaluation
- **Key contributions / Limitations**

Key Contributions

- First model performing large-scale tasks on multiple domains.
- Sets blueprint for potential future AI (broadly applicable)
- Designs multi-modal architecture with blocks from diverse modalities
- Demonstrates transfer learning across domains

Limitations

- Comparison with SOTA - last few percentages, when models approach 100% is the most crucial part
- Incomplete Experimentation - Hyperparameters not tuned
- Incomplete Results Reported - Only for some tasks
- Could be less robust to adversarial samples attack

References

- <https://venturebeat.com/2017/06/19/google-advances-ai-with-one-model-to-learn-them-all/>
- <https://aidangomez.ca/multitask.pdf>
- <https://blog.acolyer.org/2018/01/12/one-model-to-learn-them-all/>
- Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.
- Chollet, François. "Xception: Deep learning with depthwise separable convolutions." arXiv preprint (2016).
- Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538 (2017).

Thank You!

Modality Nets

Image Modality Net - analogous to Xception entry flow, uses residual convolution blocks

Categorical Modality Net - analogous to Xception exit flow, Global average pooling after conv layers

Audio Modality Net - similar to Image Modality Net