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?

Apple

Text Modality

Draw?

Speak?

/ˈæpəl/

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

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

Category 127 (Male Human)

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

“To English”

“To Category”

“To French”

“To German”

“To Parse”

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

“The above represents a triumph of either apathy or civility”

“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 $\leftrightarrow$ previous outputs
- **Decoder**: decodes (input + mixture) $\rightarrow$ output representation
- **Modality Net**: unified representation $\rightarrow$ domain-specific output
MultiModel: Input $\rightarrow$ Output

Modality Nets

Input

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
r
al]
➢ Accounts for rare words

**Input Net** - \( \text{LanguageModality}_{in}(x, W_E) = W_E \cdot x \)

**Output Net** - \( \text{LanguageModality}_{out}(x, W_S) = \text{Softmax}(W_S \cdot x) \)

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

\[ ConvStep_{d,s,f}(W, x) = LN(SepConv_{d,s,f}(W, ReLU(x))) \].

**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](#).
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.
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**

<table>
<thead>
<tr>
<th>Problem</th>
<th>MultiModel (joint 8-problem)</th>
<th>State of the art</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet (top-5 accuracy)</td>
<td>86%</td>
<td>95%</td>
</tr>
<tr>
<td>WMT EN → DE (BLEU)</td>
<td>21.2</td>
<td>26.0</td>
</tr>
<tr>
<td>WMT EN → FR (BLEU)</td>
<td>30.5</td>
<td>40.5</td>
</tr>
</tbody>
</table>
## Results

2. Does simultaneous training help?

<table>
<thead>
<tr>
<th>Problem</th>
<th>Joint 8-problem</th>
<th></th>
<th>Single problem</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(perplexity)</td>
<td>accuracy</td>
<td>log(perplexity)</td>
<td>accuracy</td>
</tr>
<tr>
<td>ImageNet</td>
<td>1.7</td>
<td>66%</td>
<td>1.6</td>
<td>67%</td>
</tr>
<tr>
<td>WMT EN→DE</td>
<td>1.4</td>
<td>72%</td>
<td>1.4</td>
<td>71%</td>
</tr>
<tr>
<td>WSJ speech</td>
<td>4.4</td>
<td>41%</td>
<td>5.7</td>
<td>23%</td>
</tr>
<tr>
<td>Parsing</td>
<td>0.15</td>
<td>98%</td>
<td>0.2</td>
<td>97%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem</th>
<th>Alone</th>
<th>W/ ImageNet</th>
<th>W/ 8 Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(ppl)</td>
<td>acc.</td>
<td>full</td>
</tr>
<tr>
<td>Parsing</td>
<td>0.20</td>
<td>97.1%</td>
<td>11.7%</td>
</tr>
</tbody>
</table>
Results

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

MoE, Attention - language experts

<table>
<thead>
<tr>
<th>Problem</th>
<th>All Blocks</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(perplexity)</td>
<td>accuracy</td>
<td>log(perplexity)</td>
<td>accuracy</td>
</tr>
<tr>
<td>ImageNet</td>
<td>1.6</td>
<td>67%</td>
<td>1.6</td>
<td>66%</td>
</tr>
<tr>
<td>WMT EN→FR</td>
<td>1.2</td>
<td>76%</td>
<td>1.3</td>
<td>74%</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>67%</td>
<td>1.4</td>
<td>72%</td>
</tr>
</tbody>
</table>
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/
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