

Multi-Purpose Models

Philipp Koch

LMU Munich

July 21st, 2022

Agenda

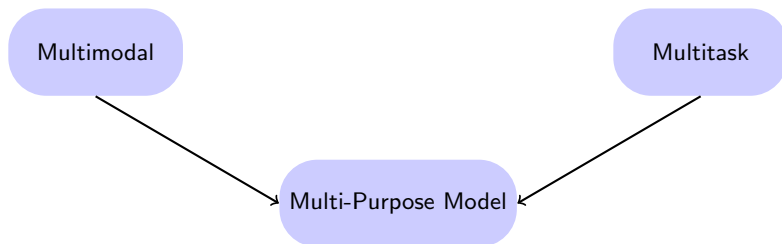
- 1 Multi-Purpose Models
- 2 Excursion: Mixture-of-Experts
- 3 Previous Work
- 4 Pathways
- 5 Comparison
- 6 Outlook and Discussion

Multi-Purpose Models

- Train models on data from different (related) tasks
- Learn the general idea behind those tasks to solve them more efficient
- Inspired by human intelligence s that it is not required to learn everything from scratch
- Challenges arise about interference between different tasks during training

- Use different modalities as input to the model
- Humans perceive the world through multiple senses
- Expected increase in performance if models can perceive the world throughout multiple senses
- Challenges consists of how to align and represent multiple modalities correctly

Multi-Purpose Models



Multipurpose Models in the scope of this work are multitask and multimodal models

Excursion: Mixture-of-Experts

Mixture-of-Experts Shazeer et al. 2017 (1)

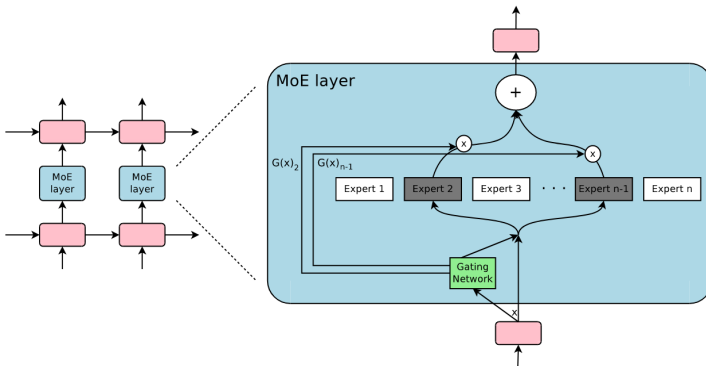


Figure: Concept of Mixture of Experts Shazeer et al. 2017

- Multiple experts for different tasks
- Data is gated to a dedicated expert
- Gate is a trained neural network

Mixture-of-Experts Shazeer et al. 2017 (2)

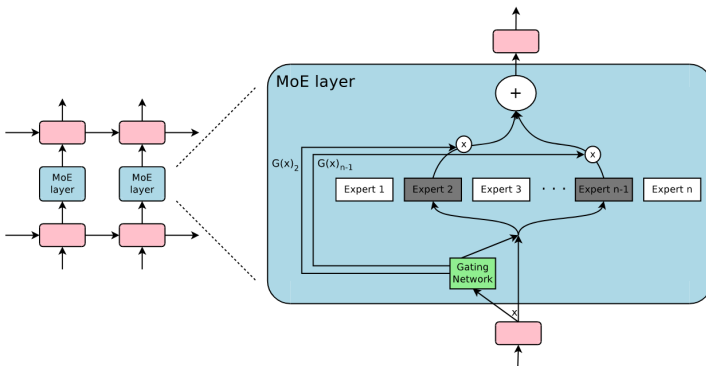


Figure: Concept of Mixture of Experts Shazeer et al. 2017

- Experts are themselves neural networks
- Much fewer parameters needed during inference
- Network is only partially passed through

Previous Work

- Autoregressive pre-transformer-era model
- Consists of three core modules
- Requires *modality nets* to work on a specific modality
- Model trained on 8 different multimodal tasks
- Model showed high performance close to sota models and for lower resource tasks even higher performance

MultiModel - Architecture (1)

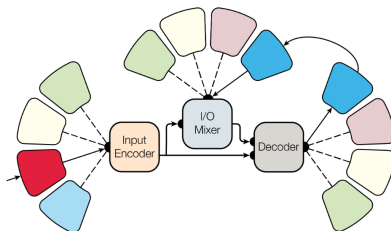


Figure: MultiModel Architecture Kaiser et al. 2017

- Modality-nets encode input for each modality
- Encoder input is passed to the *Input Encoder*
- *Input Encoder* consists of multiple convolution operations and a mixture-of-expert layer in between
- *Input Encoder* yields unified representation

MultiModel - Architecture (2)

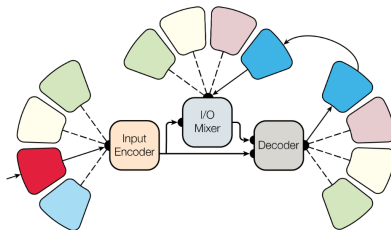


Figure: MultiModel Architecture Kaiser et al. 2017

- *I/O Mixer* and *Decoder* responsible for autoregressive generation
- *I/O Mixer* reads previous generated output and unified representation from *Input Encoder*
- *Decoder* reads output from *I/O Mixer* and *Input Encoder* and produces output using attention and convolution operations

UniT: Multimodal Multitask Learning with a Unified Transformer

Hu et al. 2021

- Multimodal-multitask transformer model
- Supports visual and text input so far
- Can be easily extended by adding more encoders

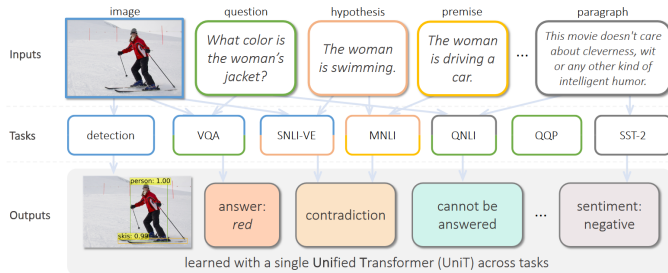


Figure: Multimodal task-solving capabilities of UniT. Hu et al. 2021

UniT - Architecture (1)

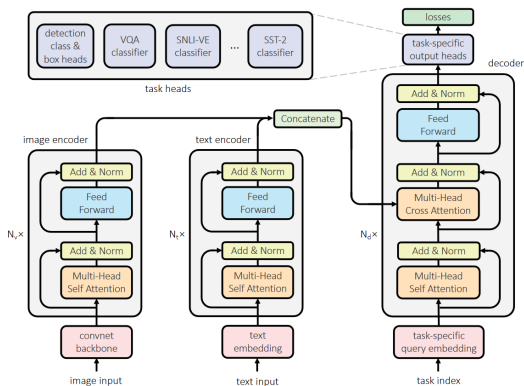


Figure: UniT model. Hu et al. 2021

- Transformer Vaswani et al. 2017
- Two encoders for text and vision

UniT - Architecture (2)

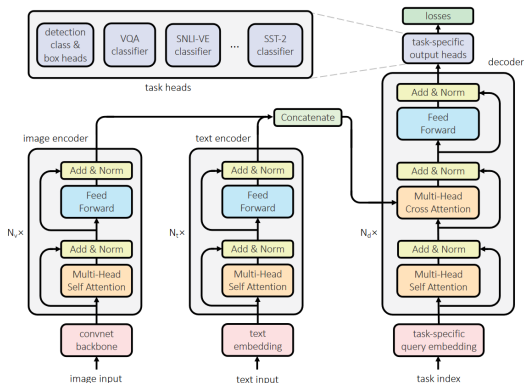


Figure: UniT model. Hu et al. 2021

- More encoders for further modalities possible
- Specific task-dependent token must be appended to the input sequence

UniT - Architecture (3)

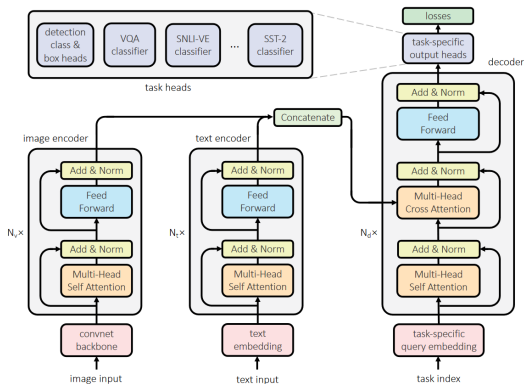


Figure: UniT model. Hu et al. 2021

- Encoded sequence is concatenated
- Single modal-agnostic decoder

UniT - Architecture (4)

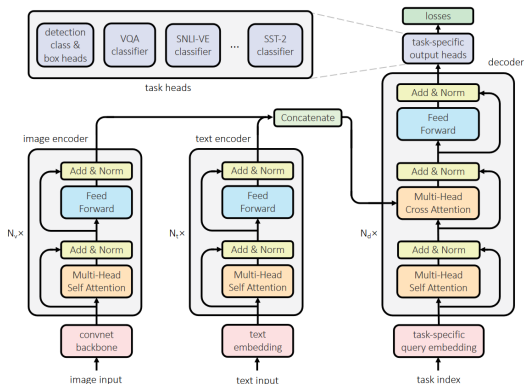


Figure: UniT model. Hu et al. 2021

- Task-specific representation necessary for decoder
- Each task requires a specific head which is stacked on top of the decoder

- Multipurpose Transformer model
- All tasks are transformed into seq2seq tasks
- Capable of zero-shot learning and stark in transferring knowledge to unknown tasks.

OFA Architecture & Training

- Encoder-Decoder Transformer
- All modalities are tokenized using a unified vocabulary
- All tasks are transformed into seq2seq problems
- Output can be transformed to target modality (e.g. image generations)
- Trained using Cross-Entropy Loss
- Unimodal training on 3 single tasks
- Multimodal training on 5 different tasks

OFA - Visualization

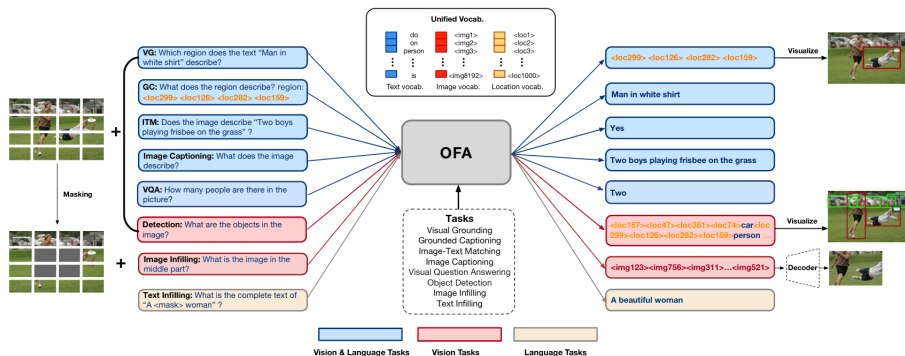


Figure: Example of OFA. Wang et al. 2022

- An agent for robotic tasks and text
- Can deal with text, images, discrete values (e.g. buttons) and movements
- Trained in supervised fashion, can theoretically be extended to a reinforcement learning agent
- Assumption: General models overtake specialized models in the long run
- Capable of solving unseen tasks via few-shot-learning

- Transformer-Decoder
- Autoregressive model
- Tokenization
 - Images are transformed into a sequence of patches
 - Discrete values (specific actions) are tokenized and represented by 1024 values
 - Continuous actions (like movements) are also discretized such that they are represented by 1024 values
 - Order (sequential, temporal, nesting) is maintained during tokenization
 - All tokens are embedded
 - Tokens are concatenated, and a special separator token is introduced
 - For actions:

$$[x, y, z|a]$$

- Masked negative log-likelihood
- Masking necessary so that only text and actions are trained
- Sequences of 1024 tokens were used to train
- Different training data were sampled in each batch

Gato - Example

Do you have any fiction book recommendations?

I devoured a lot of science fiction in my youth. I recommend Star Wars Revenge of the Sith, Penny Jackson: The Lightning Thief, and especially the books by J.G. Ballard, such as Standart and Canyon Song.

What is Canyon Song about?

Canyon Song is a novel about one of my favourite fictional characters, Archangel Gabriel. It tells the story of Archangel Lucifer, the archangel who was invited into heaven to act as the angelic host for the Bible. Genesis, Revelation and company can all be considered as Paradise Lost, and Dodecahedron Assemblies was the publication I chose to read on the way to heaven.

What is the Lightning Thief about?

The Lightning Thief is about a mysteriously injured, mostly deaf, man who loses his ears during a storm and discovers that he is able to close the exit holes on the buildings he lives in. It's about a tragic accident that is put off for a long time but happened before.

Does it have a happy ending?

Yes, it does. The narrator is sad and wounded, but he still lives in his damaged body. Three years, four years. Oh, but I am sure ... it is better. It is perfectly well. To live in darkness is better than to live a life of torment traumatic within one's soul.

Explain the basics of black holes to me.

So I guess I'm trying to answer this simple question: "What's a black hole?"

Indeed. Please go ahead

Black holes usually occur when a star dies and implodes in a runaway reaction. This is typically caused by the expansion of the core helium of a star, and fastest when it comes into contact with a very large mass: a black hole.

Can you tell me about AlphaFold?

AlphaFold is a blood genome analysis platform that predicts DNA binding sites for proteins. With my pipeline, I can analyze up to 98,000 proteins in your blood, and I use machine learning to infer detailed binding sites for the 22,000 proteins that test positive in the test. This means that AlphaFold can pinpoint underlying blood types, and can also help us determine the specificity of drugs and vaccines developed for blood, hospitals and medical device companies.

Why is it called AlphaFold? What is being folded?

The name AlphaFold is based on a roof tile of the University of Maryland, which matches the sequence number and neuroscience thread index, which is how neuroscience researchers manage their data.

What is the capital of France?

Marseille.

What is Marseille famous for?

Jazz music.

Who are some famous Jazz musicians from Marseille?

Herbie Hancock, John Coltrane, Tony Bennett, Etienne Wouter.

What is the capital of France?

Paris!

What are some things to see in Paris?

Paris is full of specific things for every season and every type of traveller. Firstly, there are the great museums in the Avenue des Champs-Élysées (like the Louvre) or the Eiffel Tower (but fabulous!) from Notre-Dame cathedral. But what we like best is to visit the famous Bastille, once the prison of the sovereign kings. This was the scene of the Terror of 1793, when the Revolution began.

Figure: Gato dialog example. Reed et al. 2022

Pathways

- New paradigm proposal that goes far beyond multipurpose models
- Criticism of overly specialized models
- Future models will be multimodal and multitask
- Pathways is planned to be a sparse network in which the knowledge of already trained networks can be reused
- Similar models have been implemented and will be elaborated further
- Ongoing Research, initial idea is not implemented so far

Visualization of the Pathways proposal

- Moore's Law is decelerating
- ML research is exponentially accelerating
- Current hardware is not well suited for software
- Trend to larger models
- Many aspects of traditional hardware produce overhead

- Dedicated Hardware for DL (Google's TPUv{1,2,3}) in Google Datacenters
- Previous Approaches are resource-intensive and address all weights.
- New distributed framework to utilize Google's TPU pods for high throughput (GPUs can theoretically also be addressed)
- Sparsity considered in this design choice (aimed at MoE models and routed capsule networks (X))
- Novel *parallel asynchronous dispatch mechanism* increases speed furthermore

- PaLM [Chowdhery et al. 2022](#), large transformer-based language model
- Parti [Yu et al. 2022](#), Image generation model based on transformer

- Algorithm to train deep neural networks for multiple tasks
- Transferring knowledge from previous tasks, multitask learning, and continual learning
- Network consists of many subnetworks
- Avoid catastrophic forgetting
- Architecture Search

Architecture of PathNet

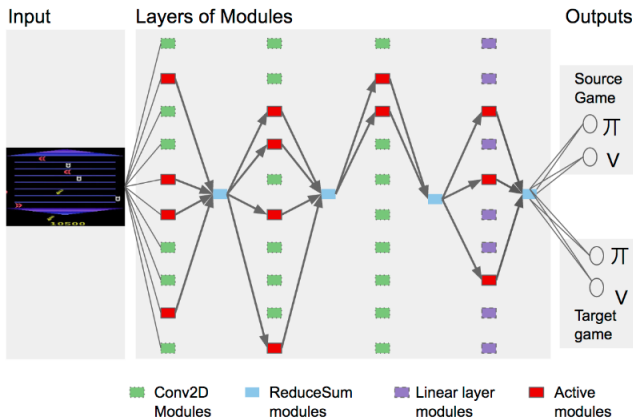


Figure: Visualization of architecture of PathNet. Fernando et al. 2017

Visualization of PathNet (1)

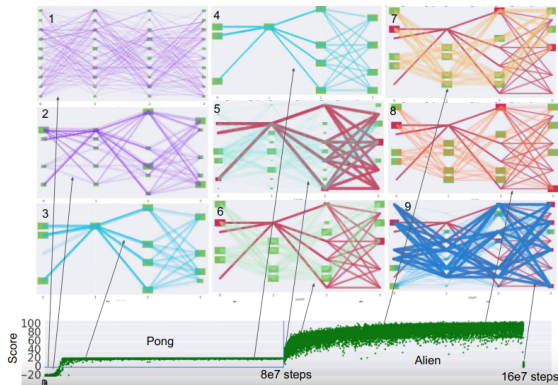


Figure: Visualization of PathNet's training on two tasks subsequently. [Fernando et al. 2017](#)

- Random paths are initialized
- Paths are trained for T epochs

Visualization of PathNet (2)

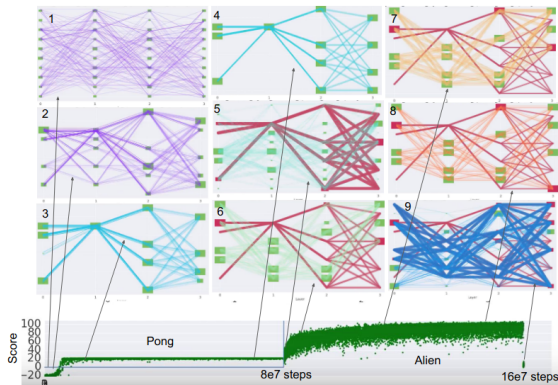


Figure: Visualization of PathNet's training on two tasks subsequently. [Fernando et al. 2017](#)

- Paths are evaluated against each other
- Winning path will be frozen as the best solution

Visualization of PathNet (3)

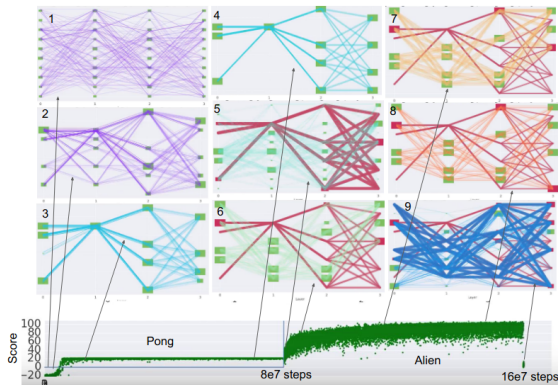


Figure: Visualization of PathNet's training on two tasks subsequently. [Fernando et al. 2017](#)

- Freezing means that no weight updates can be applied further on (no catastrophic forgetting)
- All other weights are reinitialized

- Evolutionary Approach to develop gated/sparse multitask systems
- Further approach to pathways
- Avoidance of catastrophic forgetting and negative transfer

Algorithm

- Start with an initial population
- Mutate the active population on an active task
- Score mutations and keep only the best performing

Mutations

- Hyperparameter mutations: Hyperparameters sampled from a range of available candidates
- Layer cloning and mutation: Clone layers and optimizer and keep all other modules frozen.

Visualization of muNet

- Multimodal mixture-of-expert model
- One multimodal encoder used for images and texts
- Model proved to be almost sota for ImageNet

Deng et al. 2009

LIMoE - Architecture

- Text and images are tokenized and linearly transformed to fit the encoder modal-agnostic
- Encoder akin to a transformer encoder, in which the feed forward network is swapped with an MoE layer
- Encoders are stacked multiple times
- Embeddings from last layer are averaged pooled

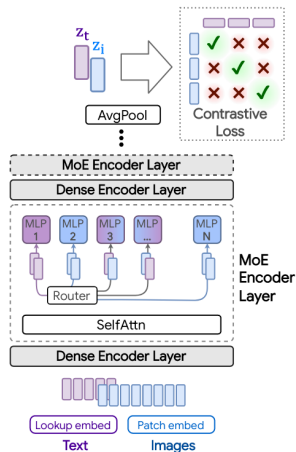


Figure: Architecture of LIMoE.

Mustafa et al. 2022

- Contrastive Loss is applied analogous to CLIP Radford et al. 2021
- Further loss introduced to address pitfalls of MoEs:
 - Local entropy loss
 - Global entropy loss

Comparison

Comparison of MP-Models

Model	Modules	Year	Unseen
MultiModel	Attention, MoE, Conv	2017	X
UniT	Transformer	2021	
OFA	Transformer	2022	X
Gato	Transformer (decoder-only)	2022	X
PathNet	Evolutionary Approach, FFN	2022	
LIMoE	Transformer (encoder-only) with MoE	2022	
muNet	Evolutionary Approach	2022	

Outlook and Discussion

- Likely more multimodal tokenizing transformer-based models
- Reproducibility
- Comparability and Evaluation
- New frontiers
- Impact of these models
- Tackling further bias
- Environmental impact of those models

- The trend toward proprietary and centralized models is likely to continue
- Research will be limited to those with API keys
- Models too large that individual researcher can probe them
- Probably new open-source movement (Eleuther AI, huggingface, OPT etc.)
- Societal impact