Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Institute for AI and Autonomous Systems

Language-Based Learning

A Short Overview of Contemporary Language Use in Robotics

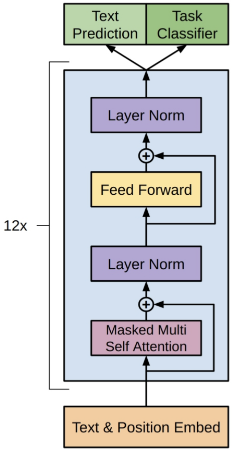
Dr. Alex Mitrevski
Master of Autonomous Systems

Structure

- (Large) Language models
- Robot learning and language

(Large) Language Models

Language Models

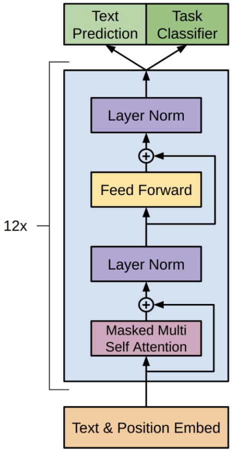


A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.

- Language models are **computational models of language** that enable language processing, understanding, and sometimes generation, to be performed

[1] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Proc. 33rd Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Language Models

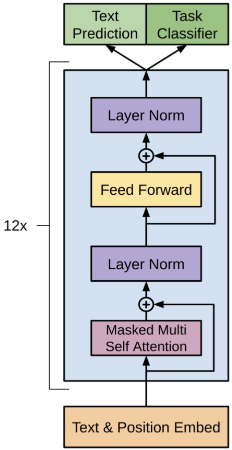


A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.

- Language models are **computational models of language** that enable language processing, understanding, and sometimes generation, to be performed
- Natural language tasks used to be performed with classical machine learning-based models; e.g. a Naive Bayes classifier could be used for text classification

[1] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Proc. 33rd Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Language Models



A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.

- Language models are **computational models of language** that enable language processing, understanding, and sometimes generation, to be performed
- Natural language tasks used to be performed with classical machine learning-based models; e.g. a Naive Bayes classifier could be used for text classification
- **Large language models** are **neural network-based language models** which have **a very large number of parameters** and which are **trained on massive datasets**
  - For instance, GPT-3 has 175 billion parameters[1]

[1] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Proc. 33rd Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tokenisation

- When processing bodies of text (e.g. full documents), a variety of **preprocessing steps** are performed; **language processing is performed on the resulting preprocessed representation**
  - For instance, stop words (e.g. a, the, etc.) and punctuation are typically irrelevant for language understanding tasks, so they might be removed in the preprocessing steps

Tokenisation

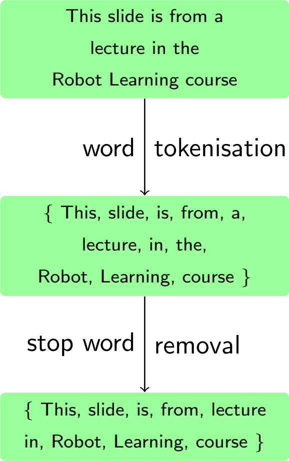
- When processing bodies of text (e.g. full documents), a variety of **preprocessing steps** are performed; **language processing is performed on the resulting preprocessed representation**
  - For instance, stop words (e.g. a, the, etc.) and punctuation are typically irrelevant for language understanding tasks, so they might be removed in the preprocessing steps
- Tokenisation is **a process of converting text into a set of constituent entities**
  - One common tokenisation strategy is to convert text into individual words — further processing is then done on the level of words

Tokenisation

- When processing bodies of text (e.g. full documents), a variety of **preprocessing steps** are performed; **language processing is performed on the resulting preprocessed representation**
  - For instance, stop words (e.g. a, the, etc.) and punctuation are typically irrelevant for language understanding tasks, so they might be removed in the preprocessing steps
- Tokenisation is **a process of converting text into a set of constituent entities**
  - One common tokenisation strategy is to convert text into individual words — further processing is then done on the level of words
- **Hierarchical token representations** can also be used
  - In this case, tokenisation can start at the level of individual characters or sub-words and progress up to words or word combinations

Tokenisation

- When processing bodies of text (e.g. full documents), a variety of **preprocessing steps** are performed; **language processing is performed on the resulting preprocessed representation**
  - For instance, stop words (e.g. a, the, etc.) and punctuation are typically irrelevant for language understanding tasks, so they might be removed in the preprocessing steps
- Tokenisation is **a process of converting text into a set of constituent entities**
  - One common tokenisation strategy is to convert text into individual words — further processing is then done on the level of words
- **Hierarchical token representations** can also be used
  - In this case, tokenisation can start at the level of individual characters or sub-words and progress up to words or word combinations



Word Embeddings

- Performing computations on language through numerical models (such as neural networks) **requires a numerical representation of tokens**
  - The bag-of-words representation or term frequency-inverse document frequency (TF-IDF) are examples of classical representations

Word Embeddings

- Performing computations on language through numerical models (such as neural networks) **requires a numerical representation of tokens**
  - The bag-of-words representation or term frequency-inverse document frequency (TF-IDF) are examples of classical representations

- A word embedding is a **vectorial token representation** that **encodes tokens in a latent space of size k**, typically produced by a neural network model
  - Embeddings are learned with respect to a vocabulary of fixed size $v >> k$
  - Inputs to embedding models are often represented as one-hot encoded vectors

Word Embeddings

- Performing computations on language through numerical models (such as neural networks) **requires a numerical representation of tokens**
  - The bag-of-words representation or term frequency-inverse document frequency (TF-IDF) are examples of classical representations

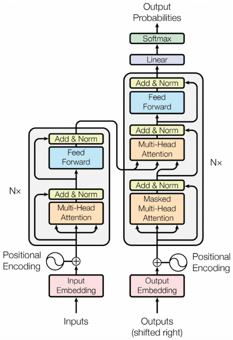
- A word embedding is a **vectorial token representation** that **encodes tokens in a latent space of size k**, typically produced by a neural network model
  - Embeddings are learned with respect to a vocabulary of fixed size $v >> k$
  - Inputs to embedding models are often represented as one-hot encoded vectors

- A variety of word embeddings have been proposed over the years — some popular ones are **word2vec, BERT,** and **ELMo**

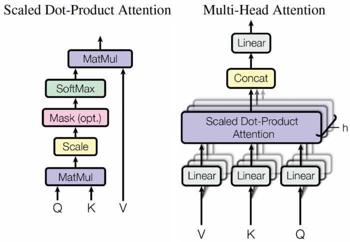
Word Embeddings

- Performing computations on language through numerical models (such as neural networks) **requires a numerical representation of tokens**
  - The bag-of-words representation or term frequency-inverse document frequency (TF-IDF) are examples of classical representations
- A word embedding is a **vectorial token representation** that **encodes tokens in a latent space of size k**, typically produced by a neural network model
  - Embeddings are learned with respect to a vocabulary of fixed size $v >> k$
  - Inputs to embedding models are often represented as one-hot encoded vectors
- A variety of word embeddings have been proposed over the years — some popular ones are **word2vec, BERT,** and **ELMo**
- A desirable feature of embedding models is that **words that have similar meanings should be close to each other in the embedding space**
  - BERT and ELMo produce context-dependent embeddings, as they are learned by considering surrounding words

Transformer



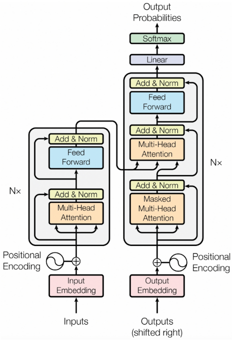
A. Vaswani et al., "Attention Is All You Need," in *31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.



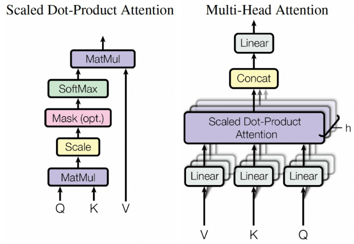
A. Vaswani et al., "Attention Is All You Need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.

- Most large language models are based on the so-called **transformer architecture**

Transformer



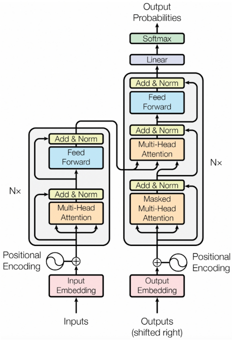
A. Vaswani et al., "Attention Is All You Need," in *31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.



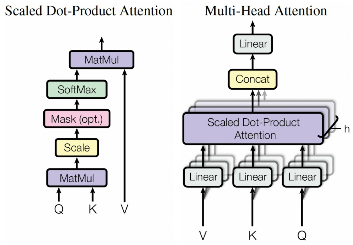
A. Vaswani et al., "Attention Is All You Need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.

- Most large language models are based on the so-called **transformer architecture**
- The main component of the transformer is an **attention layer**, which can be seen as computing token importance factors as a result of other tokens in the current context
  - The context is defined as a sequence of tokens of a predefined size

Transformer



A. Vaswani et al., "Attention Is All You Need," in *31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.



A. Vaswani et al., "Attention Is All You Need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS)*, 2017.

- Most large language models are based on the so-called **transformer architecture**
- The main component of the transformer is an **attention layer**, which can be seen as computing token importance factors as a result of other tokens in the current context
  - The context is defined as a sequence of tokens of a predefined size
- Transformer networks generally use **multi-head attention layers**, which combine the outputs of multiple individual attention layers to produce a joint attention output

Robot Learning and Language

Why Does Language Matter for Robotics?

Natural communication with people

The ability to use language for human-robot communication eliminates the need for designing specialised, less natural communication interfaces

Why Does Language Matter for Robotics?

Natural communication with people

The ability to use language for human-robot communication eliminates the need for designing specialised, less natural communication interfaces

Simplified task description

Language is an interface through which tasks — both their overall and intermediate objectives — can be described in a simple, general manner

Why Does Language Matter for Robotics?

Natural communication with people

The ability to use language for human-robot communication eliminates the need for designing specialised, less natural communication interfaces

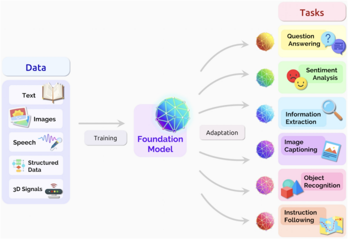
Simplified task description

Language is an interface through which tasks — both their overall and intermediate objectives — can be described in a simple, general manner

Rich data source

(Written) Language sources contain information about a variety of aspects relevant for existing in human-centred environments

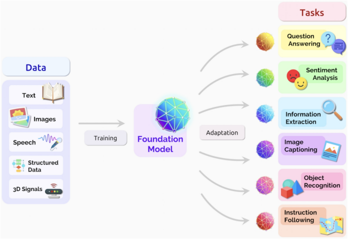
Foundation Models



R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," *CoRR*, vol. abs/2108.07258, July 2022. Available: https://arxiv.org/abs/2108.07258.

- A **foundation model** is a (neural network-based) model that is **trained on very large, diverse data**
  - Depending on the model's purpose, it can be trained on a single data modality or on multimodal data

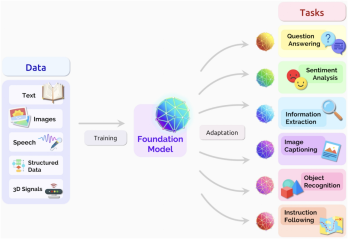
Foundation Models



R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," *CoRR*, vol. abs/2108.07258, July 2022. Available: https://arxiv.org/abs/2108.07258.

- A **foundation model** is a (neural network-based) model that is **trained on very large, diverse data**
  - Depending on the model's purpose, it can be trained on a single data modality or on multimodal data
- The main purpose of such a model is **to be used as a basis for learning specialised tasks**
  - Using a pretrained foundation model as a basis for learning another task is an example of transfer learning

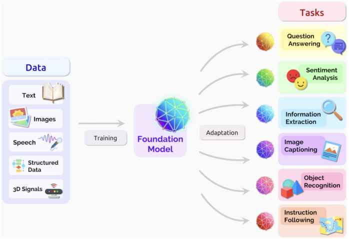
Foundation Models



R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," *CoRR*, vol. abs/2108.07258, July 2022. Available: https://arxiv.org/abs/2108.07258.

- A **foundation model** is a (neural network-based) model that is **trained on very large, diverse data**
  - Depending on the model's purpose, it can be trained on a single data modality or on multimodal data
- The main purpose of such a model is **to be used as a basis for learning specialised tasks**
  - Using a pretrained foundation model as a basis for learning another task is an example of transfer learning
- You are certainly familiar with at least one foundation model — the GPT family of models are foundation models

Foundation Models



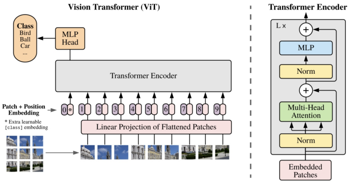
R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," *CoRR*, vol. abs/2108.07258, July 2022. Available: https://arxiv.org/abs/2108.07258.

- A **foundation model** is a (neural network-based) model that is **trained on very large, diverse data**
  - Depending on the model's purpose, it can be trained on a single data modality or on multimodal data
- The main purpose of such a model is **to be used as a basis for learning specialised tasks**
  - Using a pretrained foundation model as a basis for learning another task is an example of transfer learning
- You are certainly familiar with at least one foundation model — the GPT family of models are foundation models

"A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks..." (Bommasani et al., 2022)

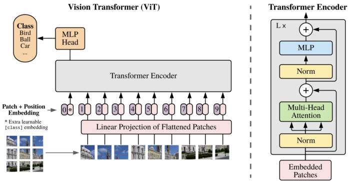
Vision Transformers

- Transformers were originally used only for language processing, but they have since been used for images as well



A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

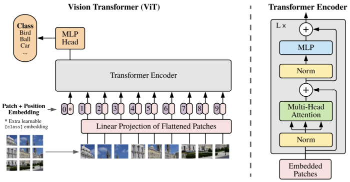
Vision Transformers



A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

- Transformers were originally used only for language processing, but they have since been used for images as well
- In a vision transformer, **an image is split into image patches** and **an embedding is computed for each individual patch**
  - The patches together with their positions are then observed as a sequence of image tokens

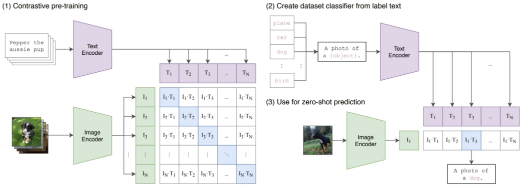
Vision Transformers



A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

- Transformers were originally used only for language processing, but they have since been used for images as well
- In a vision transformer, **an image is split into image patches** and **an embedding is computed for each individual patch**
  - The patches together with their positions are then observed as a sequence of image tokens
- **Once this "image tokenisation" is done, a transformer architecture as discussed before can be used for processing the image**
  - Attention layers use embeddings as an input, which actually makes them independent on the input modality — as long as the modality can be appropriately embedded, a transformer is applicable

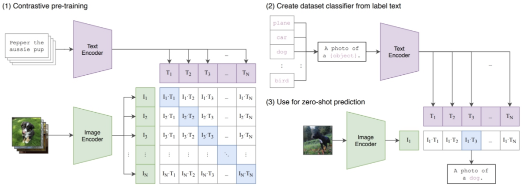
Vision-Language Models



A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Machine Learning, PMLR*, 2021, pp. 8748–8763.

- For most useful everyday tasks, **language is just an abstract representation of the world** — vision makes it possible to **ground language to real-world concepts and entities**

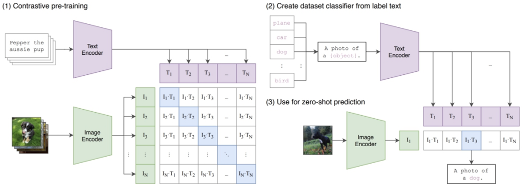
Vision-Language Models



A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Machine Learning, PMLR*, 2021, pp. 8748–8763.

- For most useful everyday tasks, **language is just an abstract representation of the world** — vision makes it possible to **ground language to real-world concepts and entities**
- A model that **combines visual and language inputs for making predictions** is referred to as a **vision-language model**

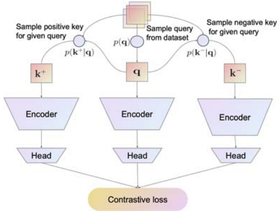
Vision-Language Models

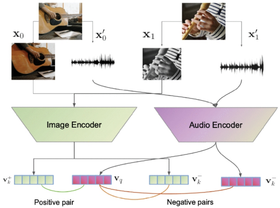


A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Machine Learning, PMLR*, 2021, pp. 8748–8763.

- For most useful everyday tasks, **language is just an abstract representation of the world** — vision makes it possible to **ground language to real-world concepts and entities**
- A model that **combines visual and language inputs for making predictions** is referred to as a **vision-language model**
- Such models are commonly learned using **contrastive learning**
  - Training requires alignment between the visual and language data

Contrastive Learning





P. H. Le-Khac, G. Healy and A. F. Smeaton, "Contrastive Representation Learning: A Framework and Review," in *IEEE Access*, vol. 8, pp. 193907-193934, 2020.

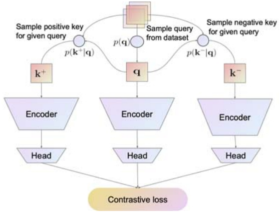
- In general, contrastive learning is concerned with **learning a distance function** $d : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ such that[2]

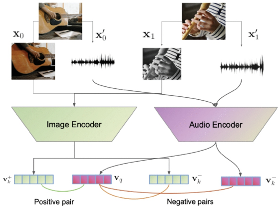
$$d(\boldsymbol{p}, \boldsymbol{p}^+) < d(\boldsymbol{p}, \boldsymbol{p}^-)$$

where $\boldsymbol{p}^+$ is a positive example and $\boldsymbol{p}^-$ is a negative example with respect to $\boldsymbol{p}$

[2] G. Chechik et al., "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

Contrastive Learning





P. H. Le-Khac, G. Healy and A. F. Smeaton, "Contrastive Representation Learning: A Framework and Review," in *IEEE Access*, vol. 8, pp. 193907-193934, 2020.

- In general, contrastive learning is concerned with **learning a distance function** $d : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ such that[2]

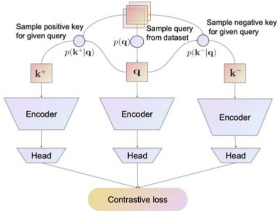
$$d(\boldsymbol{p}, \boldsymbol{p}^+) < d(\boldsymbol{p}, \boldsymbol{p}^-)$$

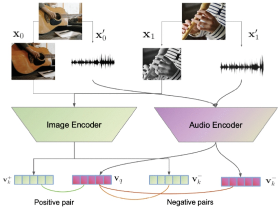
  where $\boldsymbol{p}^+$ is a positive example and $\boldsymbol{p}^-$ is a negative example with respect to $\boldsymbol{p}$

- When applied to a single modality, this objective encourages the creation of **an embedding space where similar inputs are closer to each other than dissimilar inputs**

[2] G. Chechik et al., "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

Contrastive Learning





P. H. Le-Khac, G. Healy and A. F. Smeaton, "Contrastive Representation Learning: A Framework and Review," in *IEEE Access*, vol. 8, pp. 193907-193934, 2020.

- In general, contrastive learning is concerned with **learning a distance function** $d : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ such that[2]

$$d(\boldsymbol{p}, \boldsymbol{p}^+) < d(\boldsymbol{p}, \boldsymbol{p}^-)$$

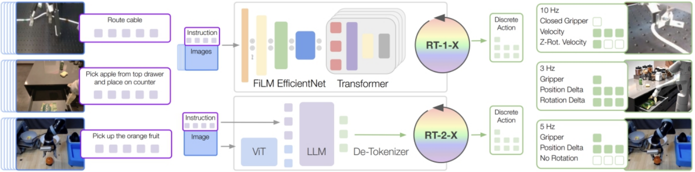
  where $\boldsymbol{p}^+$ is a positive example and $\boldsymbol{p}^-$ is a negative example with respect to $\boldsymbol{p}$

- When applied to a single modality, this objective encourages the creation of **an embedding space where similar inputs are closer to each other than dissimilar inputs**

- In the multimodal case, the objective encourages **a joint embedding space** that encourages similar entities to have similar representations across different modalities

[2] G. Chechik et al., "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

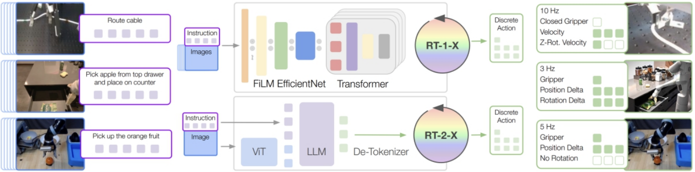
RT-X: Robot-Agnostic Foundation Models



Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models", *CoRR*, vol. abs/2310.08864, Dec. 2023. Available: https://arxiv.org/abs/2310.08864

- RT-X is a collection of very recent **foundation models trained on the X-embodiment dataset**
  - Two variants of RT-X are described, based on the recent RT-1 and RT-2 models, both of which are vision-language models
  - The outputs of both models are robot actions (represented as end effector motions and gripper opening / closing actions)

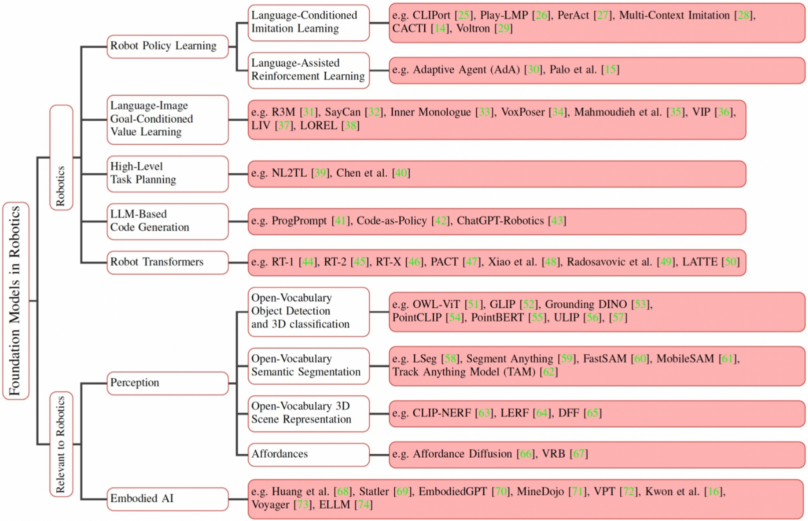
RT-X: Robot-Agnostic Foundation Models



Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models", *CoRR*, vol. abs/2310.08864, Dec. 2023. Available: https://arxiv.org/abs/2310.08864

- RT-X is a collection of very recent **foundation models trained on the X-embodiment dataset**
  - Two variants of RT-X are described, based on the recent RT-1 and RT-2 models, both of which are vision-language models
  - The outputs of both models are robot actions (represented as end effector motions and gripper opening / closing actions)
- X-embodiment combines data from multiple robots (22 in total) and a large number of robot skills (more than 500)
  - RT-X models thus aim to be foundation models applicable to different robot embodiments
  - The generalisation limitations are currently unknown though

Uses of Language / Foundation Models in Robotics



R. Firoozi et al., "Foundation Models in Robotics: Applications, Challenges, and the Future", *CoRR*, vol. abs/2312.07843, Dec. 2023. Available: https://arxiv.org/abs/2312.07843

Some Challenges with Robot Foundation Models

No safety guarantees

Current models are trained and deployed without considering safety constraints

Some Challenges with Robot Foundation Models

No safety guarantees

Current models are trained and deployed without considering safety constraints

Challenging failure analysis

The causes of failures produced by large robot models can be (mildly put) difficult to understand

Some Challenges with Robot Foundation Models

No safety guarantees

Current models are trained and deployed without considering safety constraints

Challenging failure analysis

The causes of failures produced by large robot models can be (mildly put) difficult to understand

Unknown generalisation conditions

The conditions under which generalisation between environment conditions and / or robots is possible are not well defined

Some Challenges with Robot Foundation Models

No safety guarantees

Current models are trained and deployed without considering safety constraints

Challenging failure analysis

The causes of failures produced by large robot models can be (mildly put) difficult to understand

Unknown generalisation conditions

The conditions under which generalisation between environment conditions and / or robots is possible are not well defined

Computational challenges

Robot foundation models are large and require powerful hardware to run efficiently — using them for offline execution is impossible for most robots

Summary

- Large language models are based on the transformer architecture, which includes a multitude of attention layers that operate over embedding tokens
- Vision-language models are models that are trained on aligned visual and language datasets
- Multimodal learning can be performed using contrastive learning, which results in a joint embedding space over the different modalities
- Robot foundation models, such as the recent RT-X, have been applied to various robot problems, such as task planning, policy learning, and value learning
- The general applicability of robot foundation models is conditioned on resolving various limitations with respect to safety, transparency, and efficiency