Vision, Language and Action:  
from Captioning to Embodied AI  

Lorenzo Baraldi, Marcella Cornia, Federico Landi  

{name.surname}@unimore.it  

University of Modena and Reggio Emilia, Italy
• Expertise on Sequence modelling
  • LSTM-based, Transformer-based
• Integrating Vision and Language
• Embodied AI: Vision, Language and Navigation
• Video modelling
• soon to come: Explainable AI
Image Captioning:

- Describe an image through a conditional language model
- Usually: (a set of) visual feature vectors + LSTM-based language model with attention
- One image, one caption

Show, control and tell

- Make an LSTM-based language model controllable via:
  - A sequence of regions (ordered)
  - A set (unordered)

Different captions for the same image!

• Language models take as input a sequence of regions
• Switches between one region and the next one via a learned chunk-shifting gate
  • When it’s done with the generation of chunk, it moves to the next region in the sequence

• Language models take as input a sequence of regions

• Switches between one region and the next one via a learned chunk-shifting gate
  • When it’s done with the generation of chunk, it moves to the next region in the sequence

Language models take as input a sequence of regions

Switches between one region and the next one via a learned chunk-shifting gate

- When it’s done with the generation of chunk, it moves to the next region in the sequence

• Language models take as input a sequence of regions
• Switches between one region and the next one via a learned chunk-shifting gate
  • When it’s done with the generation of chunk, it moves to the next region in the sequence

- Train on GT words and shifting gate values (obtained via NLP)
- ... and since we are at it, finetune using Reinforcement Learning
  - CIDEr wrt GT caption (as usual)
  - Plus, use the alignment between the predicted and GT chunks as reward (Needleman-Wunsh algorithm)

• What if we have an unordered set as input?
  • We can learn a network to do the sorting! $\rightarrow$ SINKHORN NETWORK
    • Approximates a derivable permutation matrix
    • Train on real data, then use the Hungarian to get the true permutation matrix.

Results when Controlling with a sequence of regions

A man sitting at a desk with a computer and a man holding a camera.
A man sitting at a desk with a computer.

Results when Controlling with a set of regions

A dog holding a frisbee in its mouth.  
A dog standing in the grass with a frisbee in its mouth.

Describing videos

• LSTMs do not show good learning capabilities on long sequences
• Plus, they do not deal with the layered structure of videos.

Our proposal

• A layered LSTM module which can adaptively modify its structure to input data.
• The result is a variable length and adaptive encoding of the video, whose length and granularity depends on the input video itself.

Latest work

- We redefine the LSTM cell, endowing it with a better connection between long and short memory.
- Solves learning issues of the peephole connection.
- Improves learning, especially on long sequences, on a variety of sequence modelling tasks.

\[ g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \]
\[ i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \]
\[ f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \]
\[ c_t = f_t \odot c_{t-1} + i_t \odot g_t \]
\[ o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \]
\[ h_t = o_t \odot \tanh(c_t) \]

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Bit per Character (BPC)</th>
<th>Fixed # Params (~ 2.2M)</th>
<th>Fixed # Hidden Units (512)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$T_{PTB}$ = 150</td>
<td>$T_{PTB}$ = 300</td>
</tr>
<tr>
<td>LSTM</td>
<td>1.334 ± 0.0006</td>
<td>1.343 ± 0.0004</td>
<td>1.386 ± 0.0005</td>
</tr>
<tr>
<td>LSTM-PH</td>
<td>1.339 ± 0.0048</td>
<td>1.343 ± 0.0009</td>
<td>1.383 ± 0.0004</td>
</tr>
<tr>
<td>LSTM-WM</td>
<td>1.299 ± 0.0005</td>
<td>1.302 ± 0.0008</td>
<td>1.299 ± 0.0005</td>
</tr>
</tbody>
</table>

Relationships between image regions are modeled via persistent memory vectors. Encoder and decoder layers are connected in a mesh-like structure.
• In our encoder, the set of keys and values is extended with learnable vectors that can encode a priori information.

• A mesh connectivity is operated through a learnable gating mechanism which modulates the contribution of each encoder layer during cross attention.

• Final goal: describing *everything* even if it is not in the training set.

• **Joint collaboration with NVAITC:**
  • Extension to multi-GPU and multi-node training
  • Optimization of data loading and processing, mixed-precision training
  • (soon to come) migration to Marconi-100

• Next research steps: extension to the description of novel objects, via self-supervised training.
Get your hands dirty

Speaksee: our library for visual-semantic tasks (captioning, retrieval, ...) built on top of PyTorch

Data-loading, Models, Evaluation

https://github.com/aimagelab/speaksee
“International Workshop on Computational Aspects of Deep Learning”

Workshop organized with NVAITC, at ICPR 2020 (January 2021)

• Discussing innovations in the field of computational optimization
• Also, a place for the NVAITC community to meet and enlarge itself!

Topics:

• Design of innovative architectures and operators for data-intensive scenarios
• Video understanding and spatio-temporal feature extraction
• Distributed reinforcement learning algorithms
• Applications of large-scale pre-training techniques
• Distributed training approaches and architectures
• HPC and massively parallel architectures in Deep Learning

See the webpage for CfP, dates and organization: www.cadl.it
Instruction:
Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.
Instruction:
Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.
Instruction:
Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.
Instruction:
Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.
Locomotion is guided by visual perception. Not only does it depend on perception, but perception depends on locomotion.

The Ecological Approach to Visual Perception
James J. Gibson

PTA is the first Transformer-like architecture merging three different modalities: text, vision, and action.

With a focus on low-level motion: a more realistic setup; closer to real-world applications

Explore and Explain — a new setting for embodied AI

The agent must jointly deal with two different tasks:

- Exploration of unseen environments
- Description of the most relevant visual features
Curiosity-driven exploration module + Transformer-based captioner

Our navigation module enriches traditional curiosity-based navigation with a penalty on repeated actions.

EX² can **identify** and **describe** relevant object during exploration

- An important means to provide some **explainability** for the actions performed by the agent
Thanks!
Questions?

Matteo Stefanini  Matteo Tomei  Federico Landi  Roberto Bigazzi  Marcella Cornia  Silvia Cascianelli  Lorenzo Baraldi  Rita Cucchiara