@DocXavi

Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giró-i-Nieto

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

Module 6

# Deep Learning for Video: Language

22nd March 2018

# Deep Learning online courses by UPC:


- [MSc course](#) (2017)
- [BSc course](#) (2018)

Next edition Autumn 2018


- [1st edition](#) (2016)
- [2nd edition](#) (2017)
- [3rd edition](#) (2018)

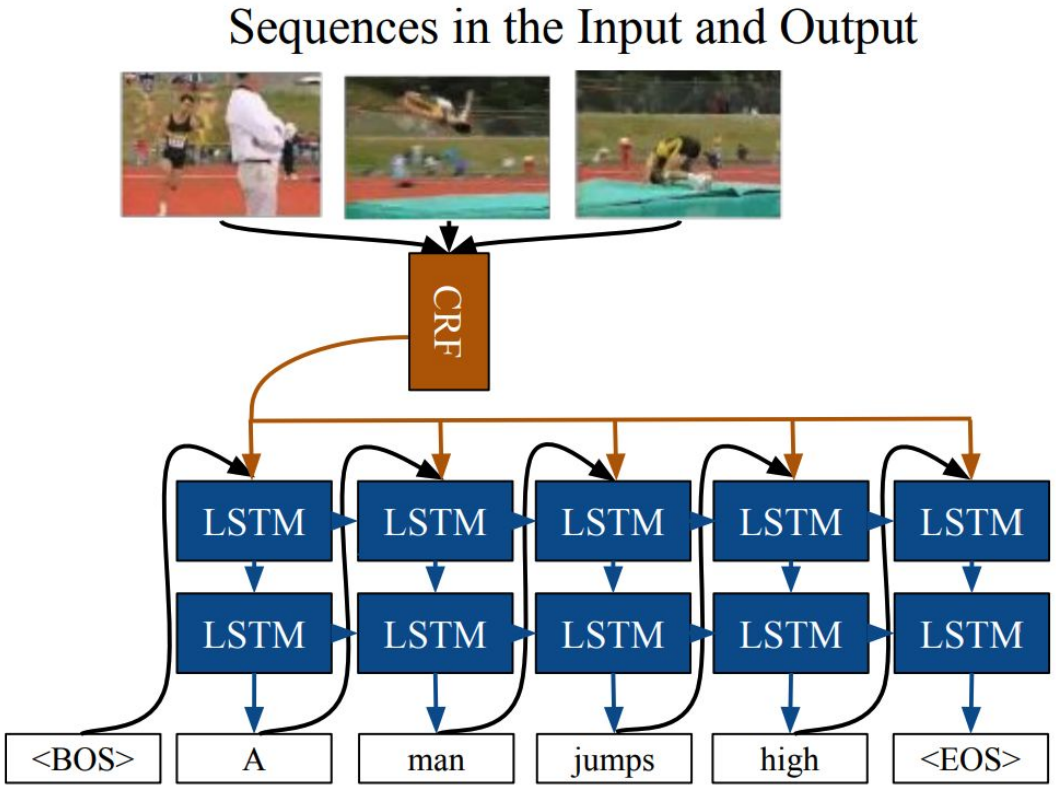
Summer School (late June 2018)


- [1st edition](#) (2017)
- [2nd edition](#) (2018)

Next edition Winter/Spring 2019

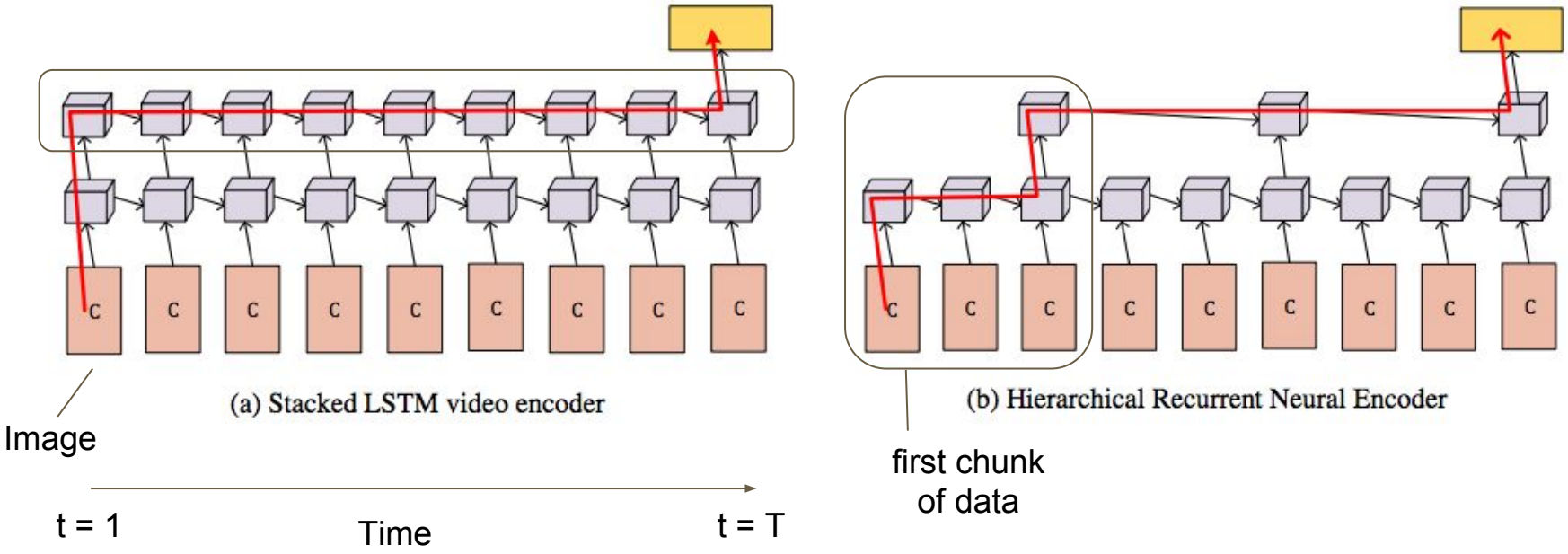

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

# Language: Captioning: RNN



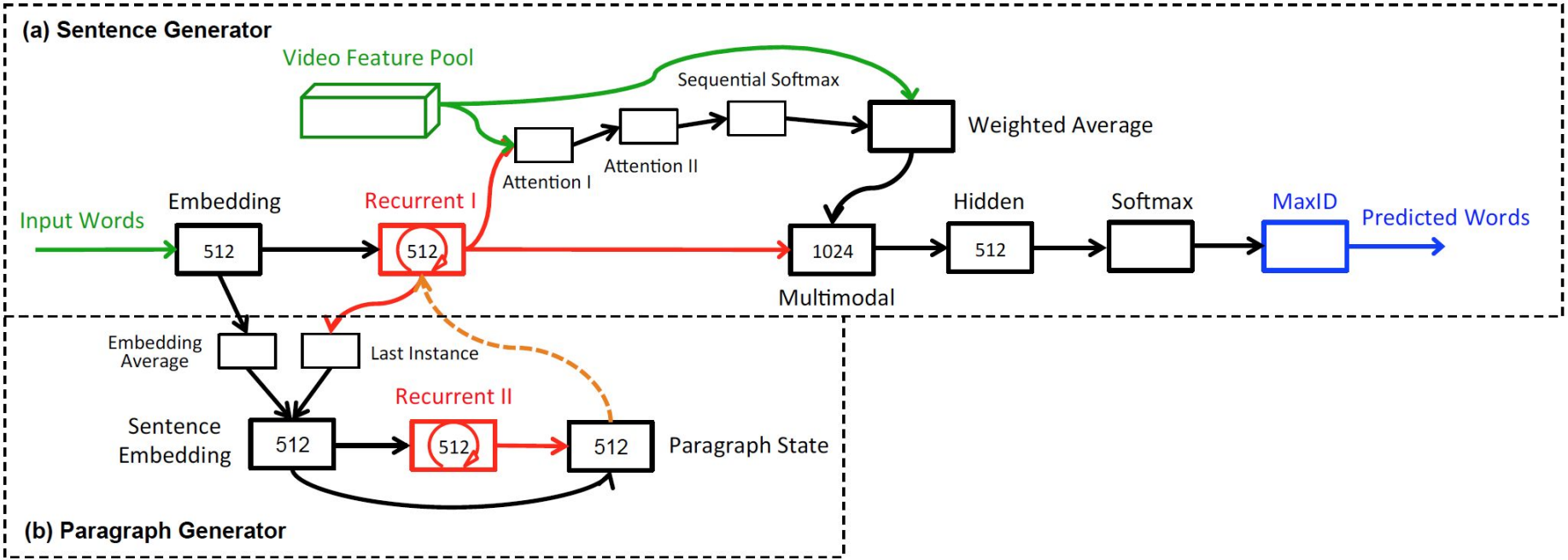

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

# Language: Captioning: Hierarchical RNN



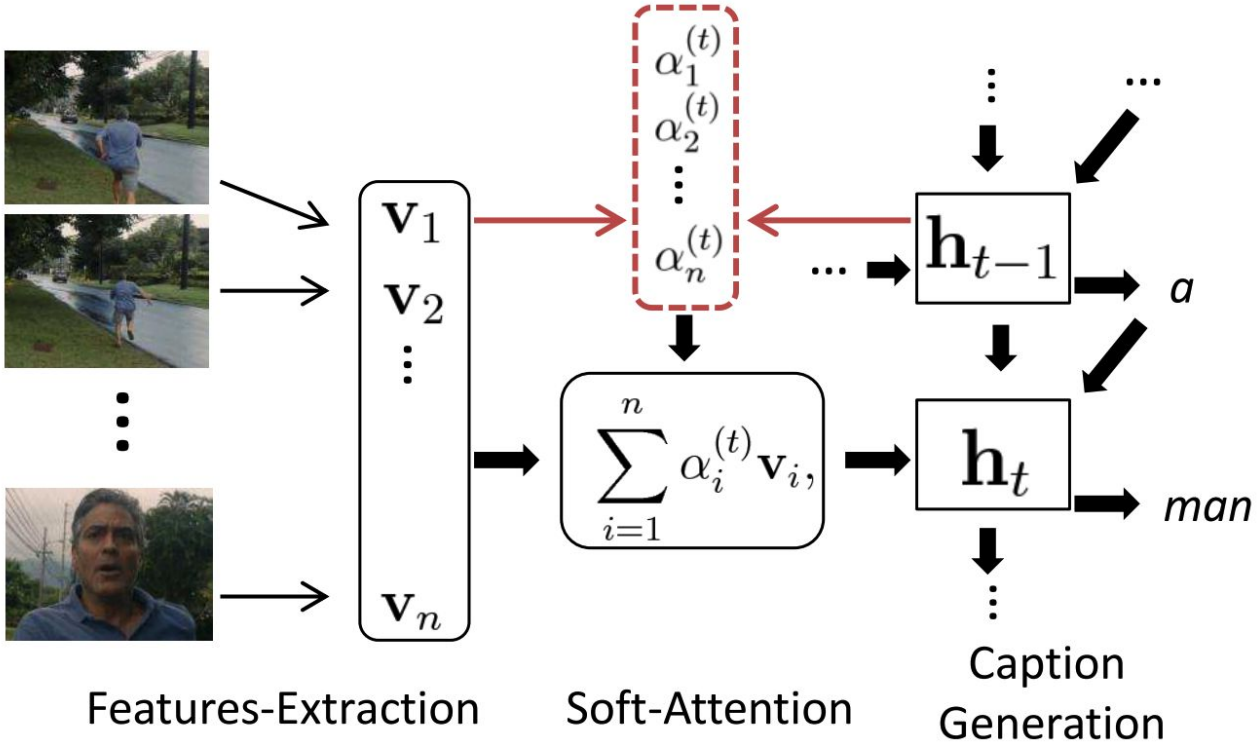

(Slides by Marc Bolaños) Pingbo Pan, Zhongwen Xu, Yi Yang,Fei Wu,Yueting Zhuang Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning, CVPR 2016.

# Captioning: Image + Hierarchical RNNS + Attention



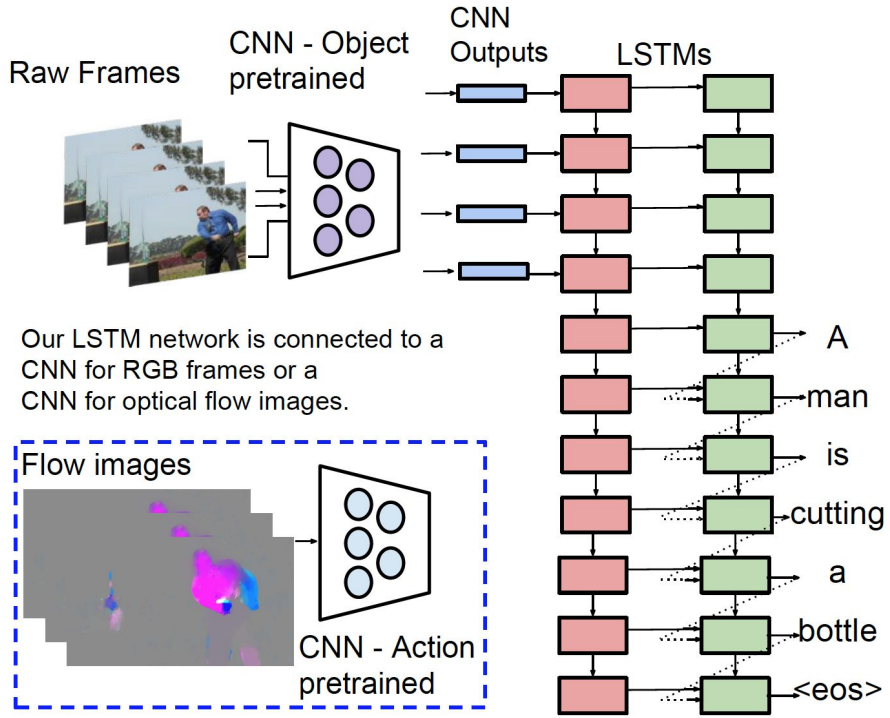

Yu, Haonan, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. "Video paragraph captioning using hierarchical recurrent neural networks." CVPR 2016.

# Video Captioning with Attention



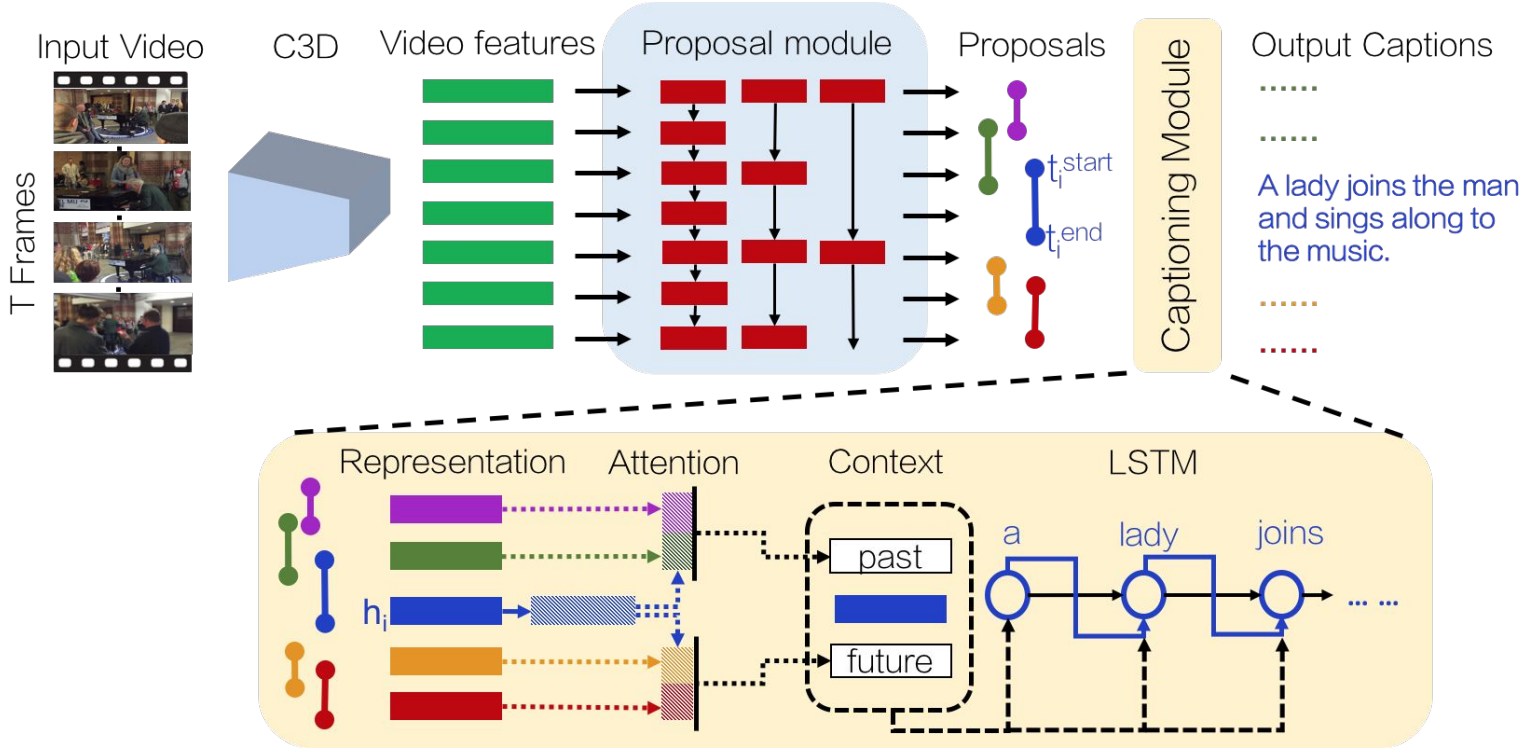

Yao, Li, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. "Describing videos by exploiting temporal structure." ICCV 2015

# Captioning: Image + Optical Flow + LSTM




Venugopalan, Subhashini, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. "Sequence to sequence-video to text." ICCV 2015

# Captioning: C3D + Proposals + LSTM




Krishna, Ranjay, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. "Dense-Captioning Events in Videos." ICCV 2017


Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." *CVPR 2017*

# Lipreading: Watch, Listen, Attend & Spell

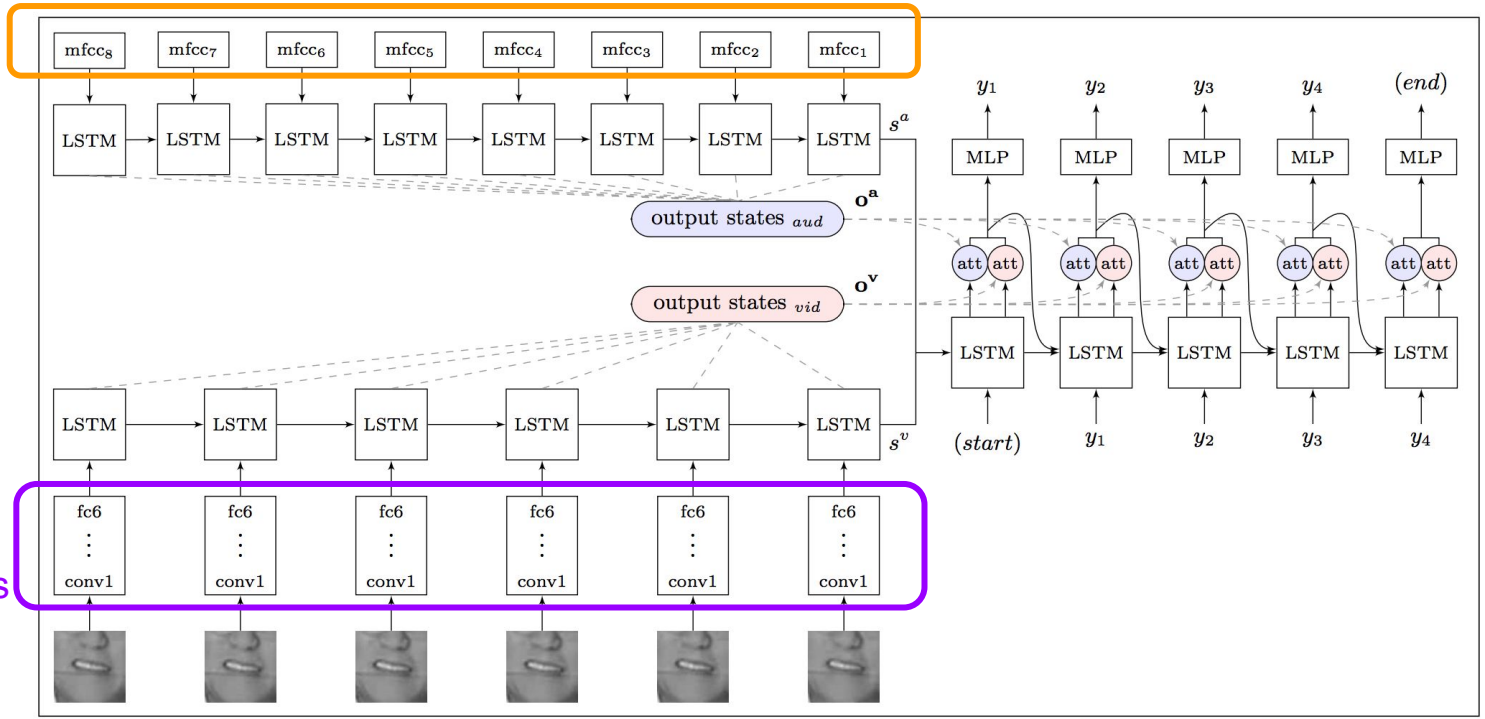



Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character $y_i$, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." *CVPR 2017*

# Lipreading: Watch, Listen, Attend & Spell

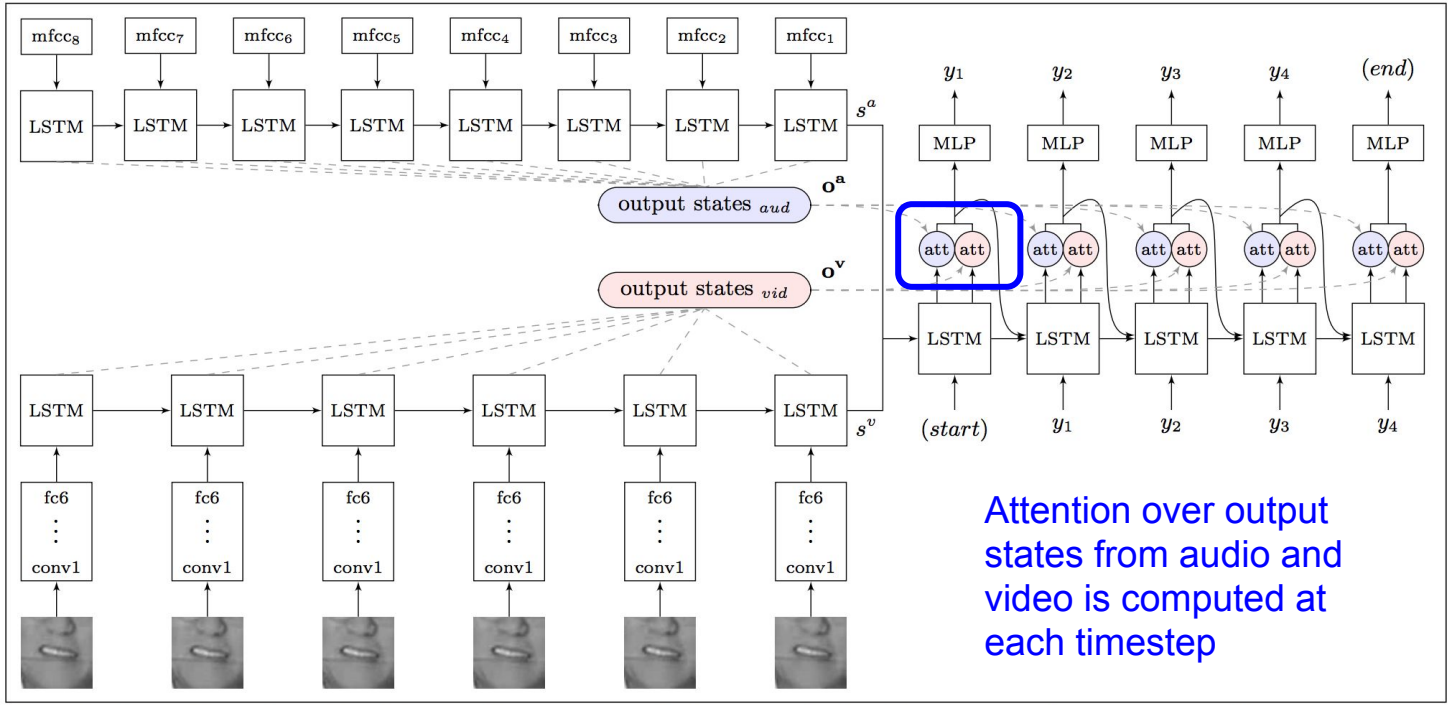
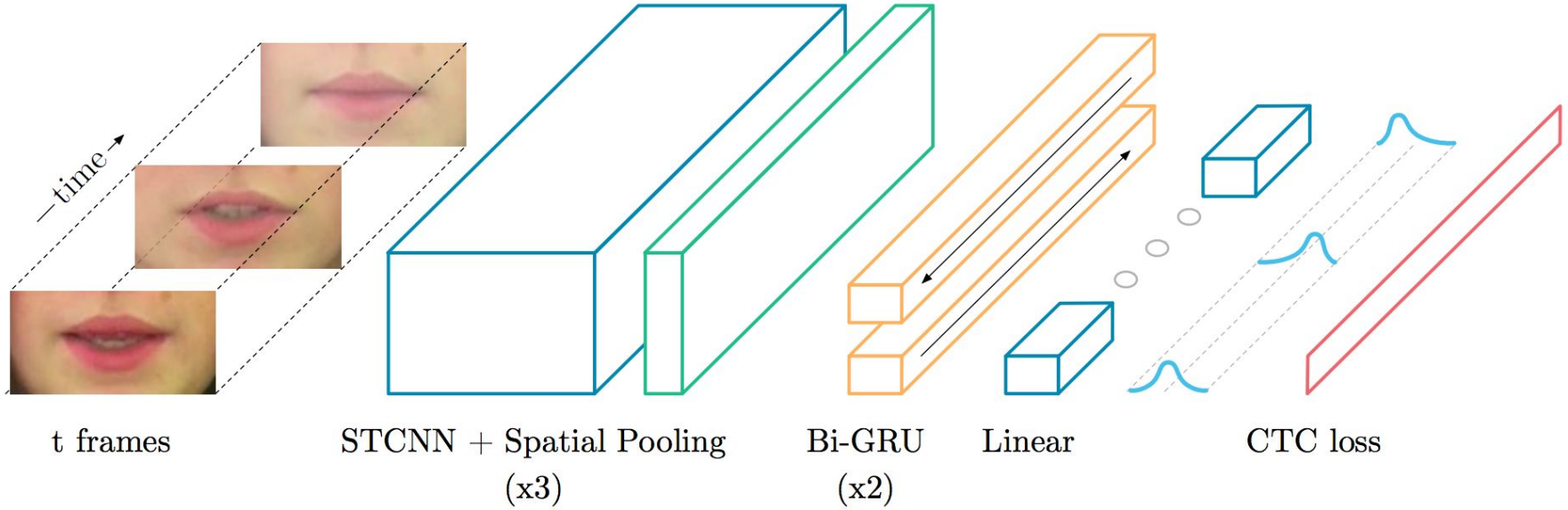



Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character $y_i$, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." *CVPR 2017*

# Lip Reading: LipNet

Input (video frames) and output (sentence) sequences are not aligned

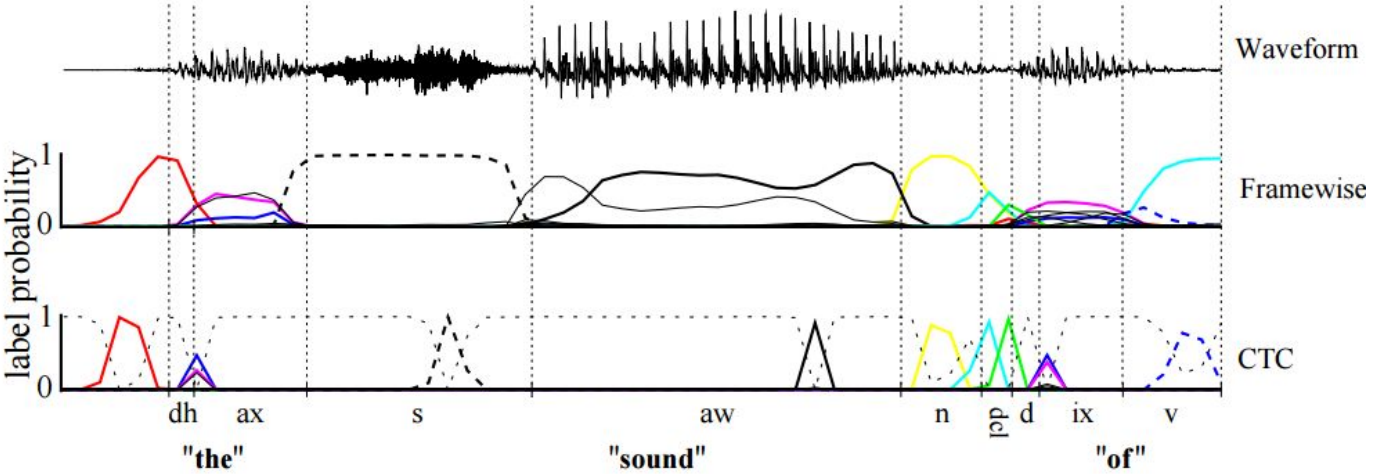



Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. "LipNet: End-to-End Sentence-level Lipreading." (2016).

# Lip Reading: LipNet

## CTC Loss: Connectionist temporal classification

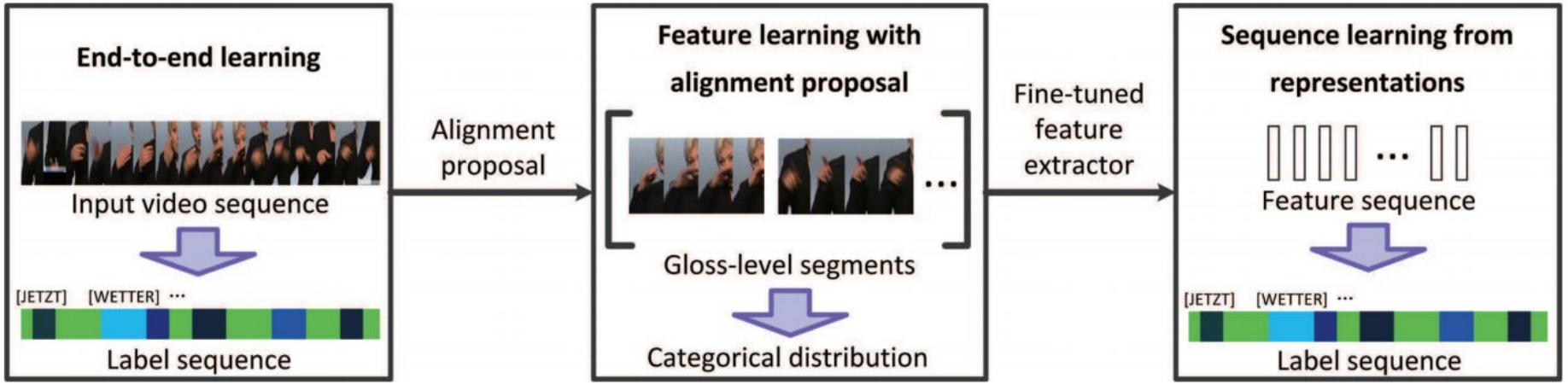
- Avoiding the need for alignment between input and output sequence by predicting an additional “_” blank word
- Before computing the loss, repeated words and blank tokens are removed
- “a _ a b _ ” == “_ a a _ _ b b” == “a a b”




Graves et al. [Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks](). ICML 2006

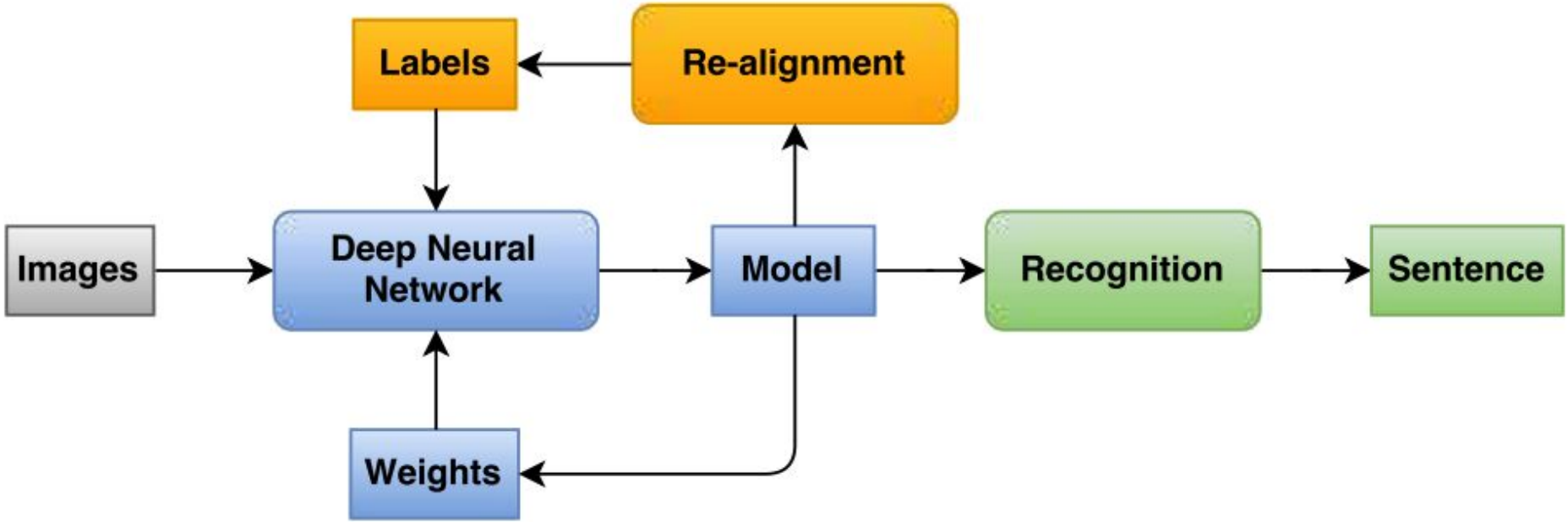

Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. "LipNet: End-to-End Sentence-level Lipreading." (2016). *[code]*

# Sign Language: RNN



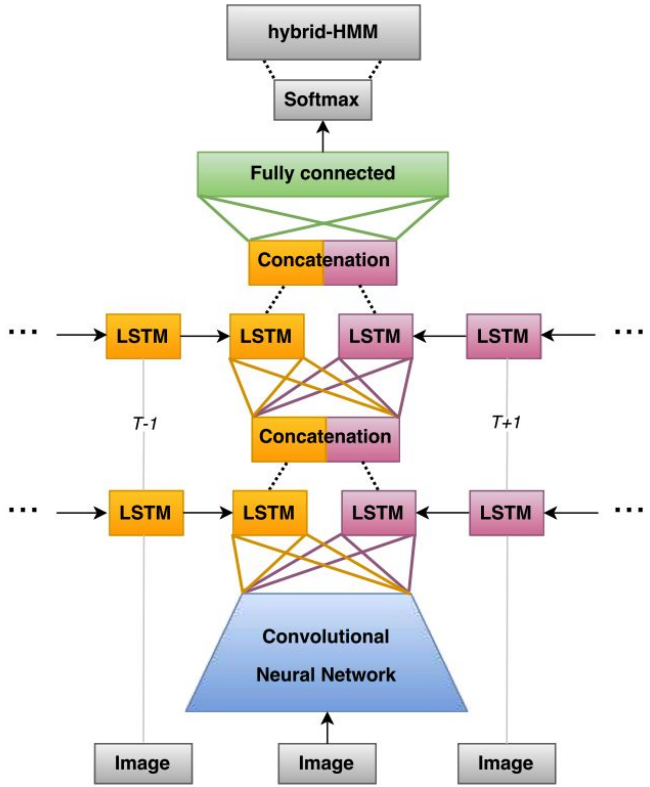

Cui, Runpeng, Hu Liu, and Changshui Zhang. "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization." CVPR 2017

# Sign Language: Re-Sign



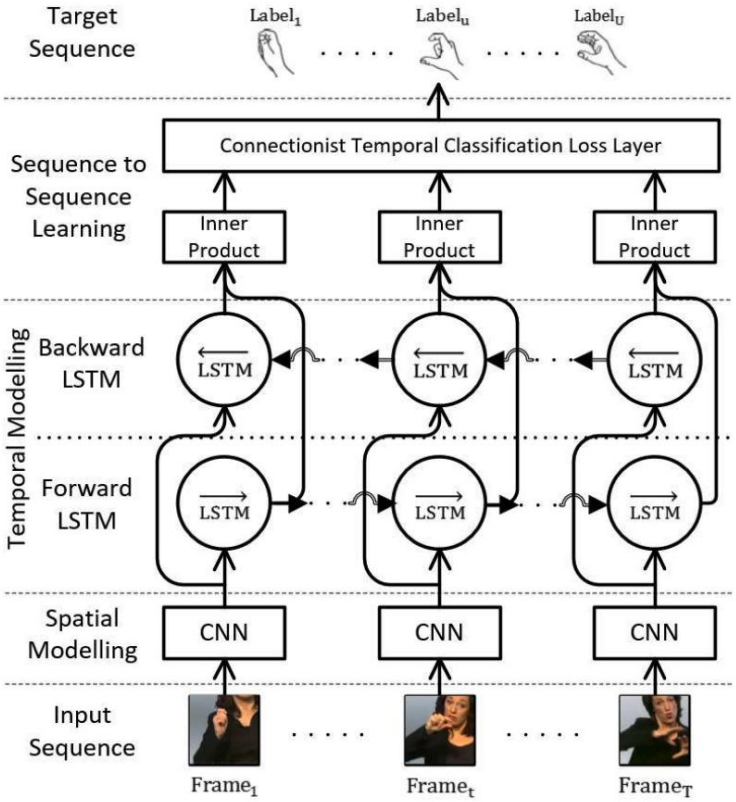

Koller, Oscar, Sepehr Zargaran, and Hermann Ney. "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs." CVPR 2017

# Sign Language: Re-Sign



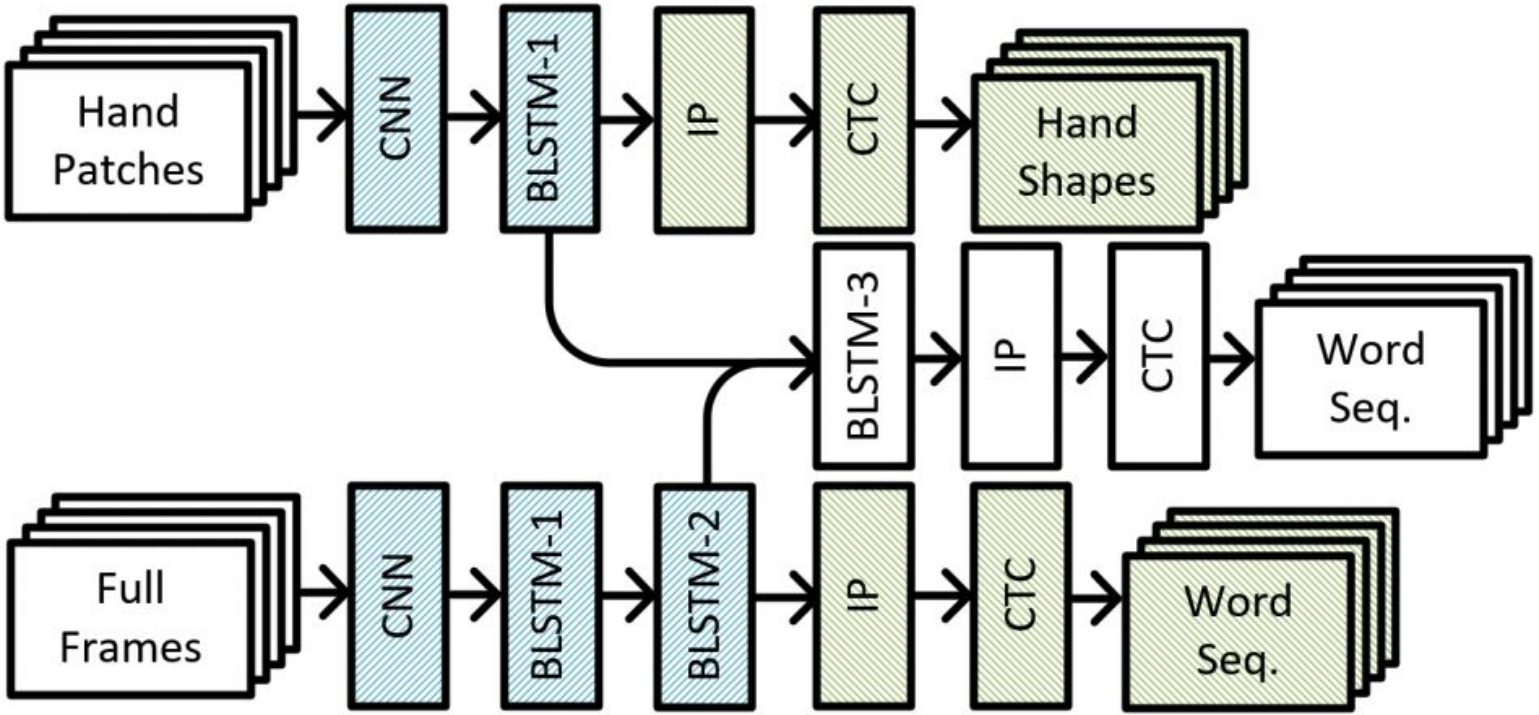

Koller, Oscar, Sepehr Zargaran, and Hermann Ney. "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs." CVPR 2017

# Sign Language: SubUNets




N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition. ICCV 2017

# Sign Language: SubUNets

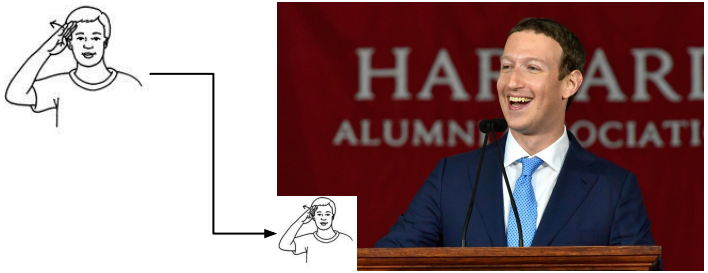



N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. SubUNets: End-to-end Hand Shape and Continuous Sign Language Recognition. ICCV 2017

# Speech2Signs (under work)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

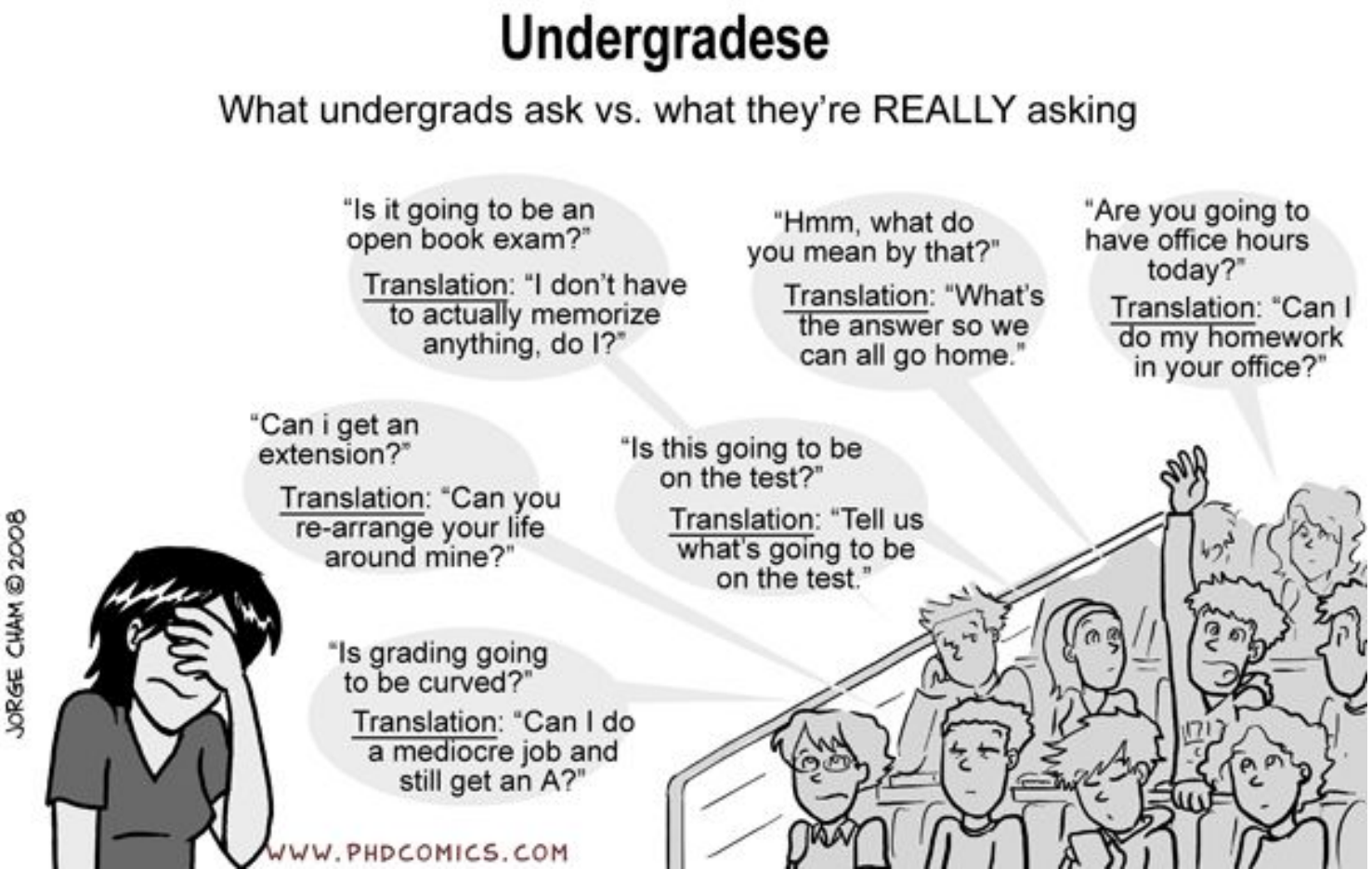
Hello → ASR → "Hello" → NMT → SLPA → Video Generator



Caffe2

facebook research

# Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM