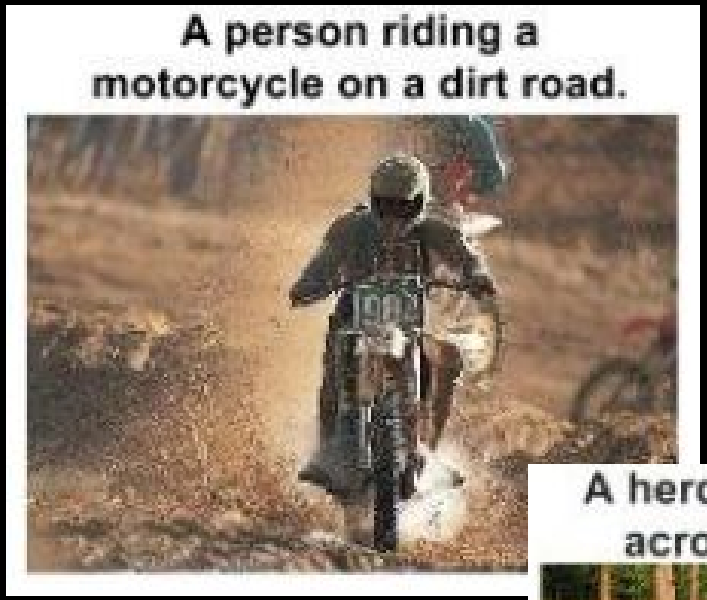
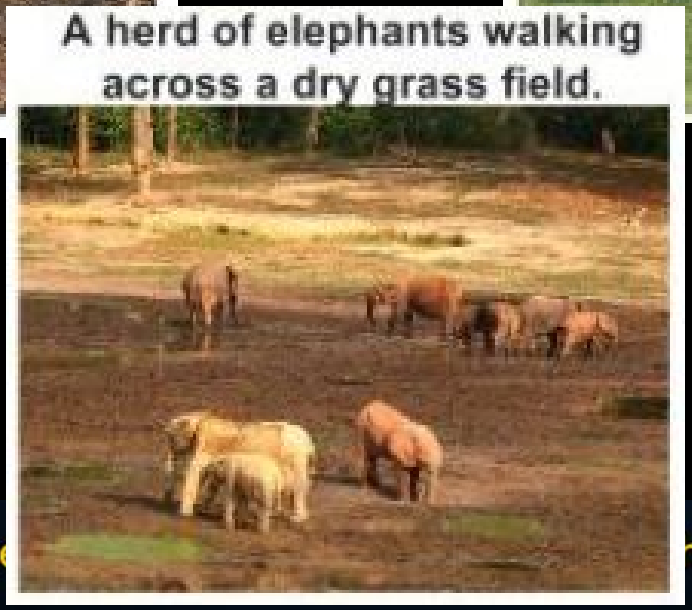
DEEP LEARNING BRASIL
SUMMER SCHOOL

Generating Image Descriptions
with Deep Learning

Otávio Calaça Xavier
otavio@inf.ufg.br

Exemplos





Vinyals, Oriol, et al. (2016).

Obrigado por ter vindo!

Otávio Calaça Xavier - otavio@inf.ufg.br

- Mestre e Doutorando em Ciência da Computação
Professor e Pesquisador no IFG e na UFG
- Consultor em Arquitetura de Software, DevOps e Machine Learning
- Pai de gêmeos

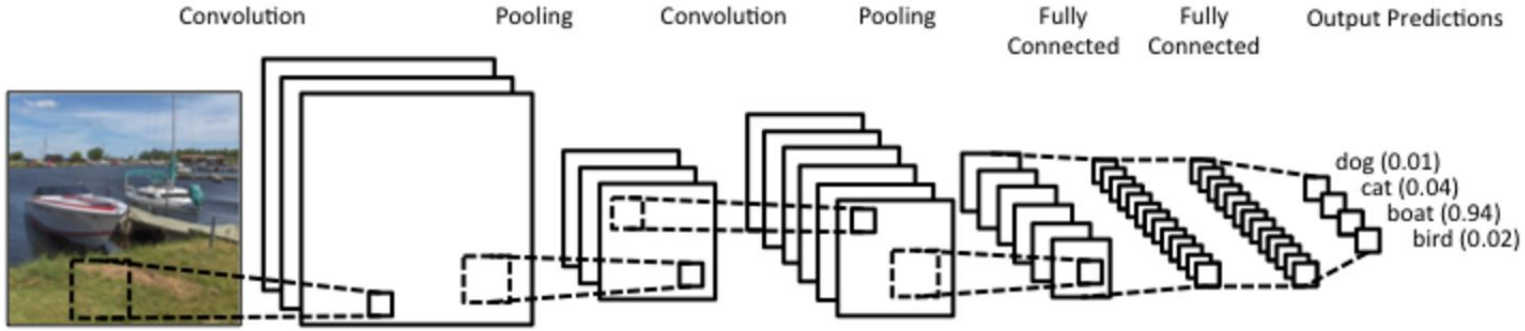
UFG UNIVERSIDADE FEDERAL DE GOIÁS

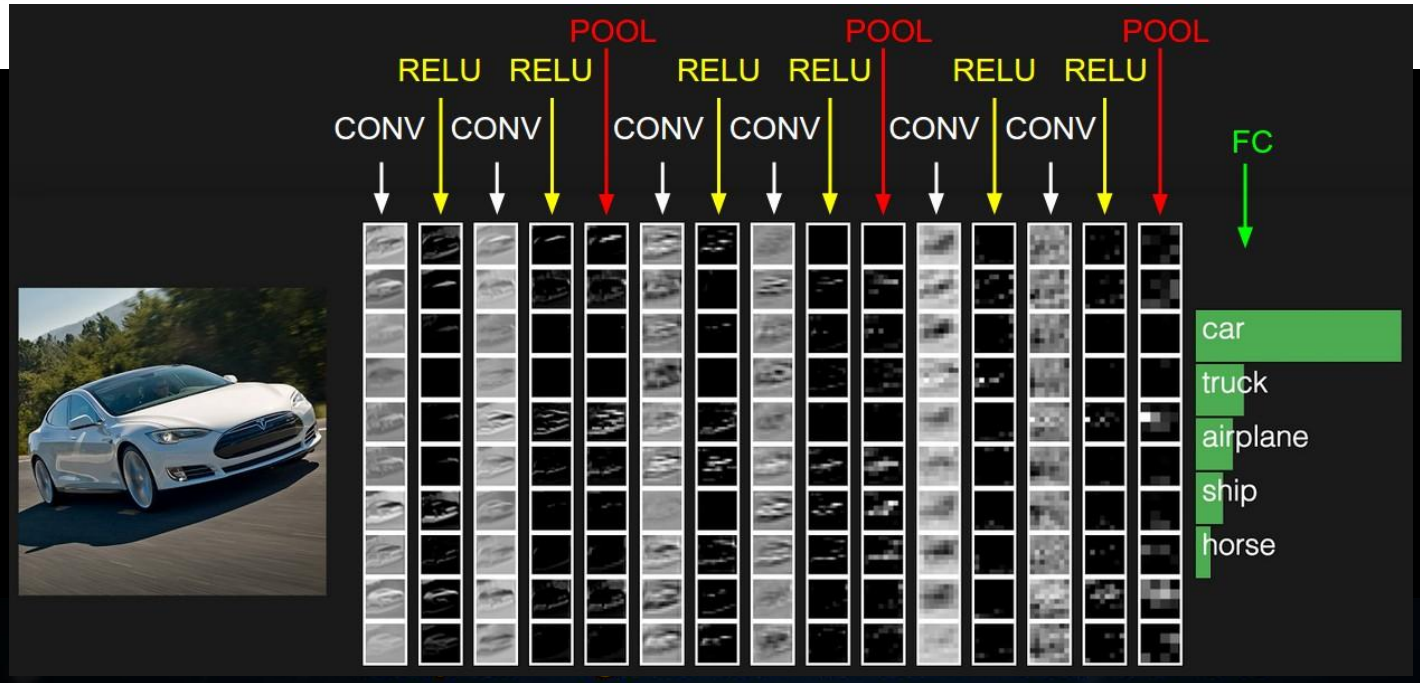
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIÁS

Agenda

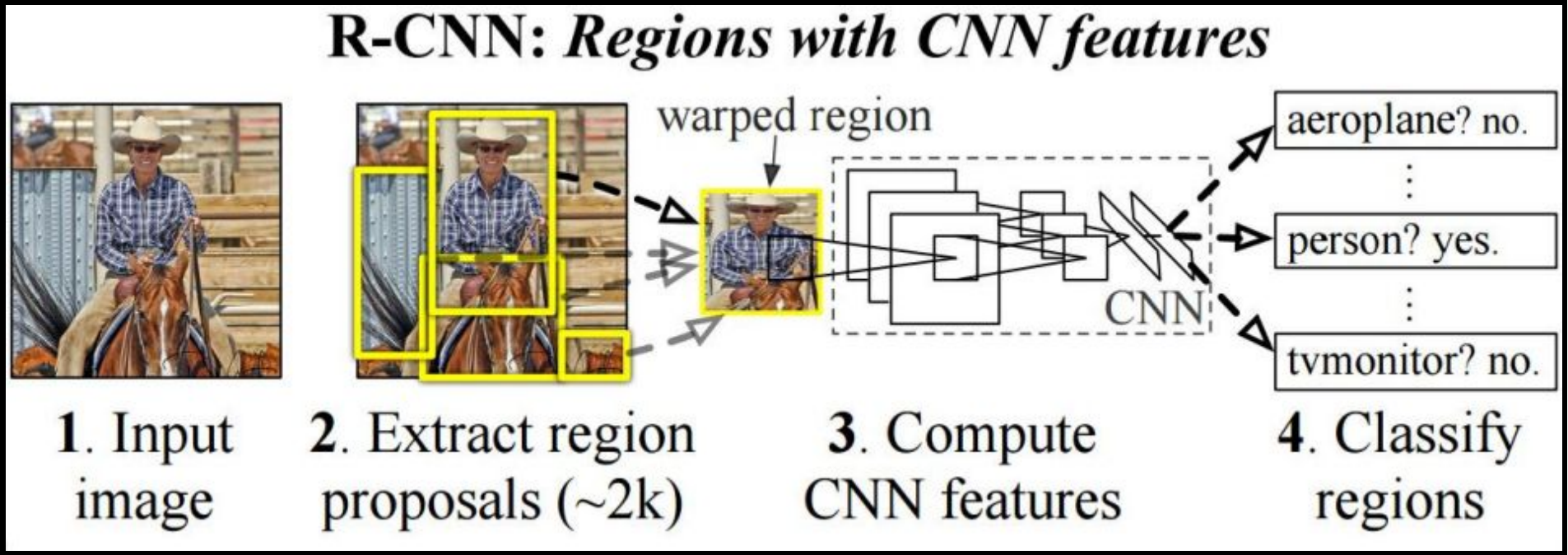
- Conceitos
  - Árvores de Dependência Tipadas.
  - Redes neurais convolucionais (CNN) e baseadas em região (R-CNN).
  - Redes Neurais Recorrentes Bidirecionais (BRNN) e Longa Memória de Curto Prazo (LSTM).
  - Modelos Baseados em Atenção.
- Evolução da Geração de Descrição de Imagens

Redes Neurais Convolucionais (CNN)





CNN baseada em região (R-CNN)

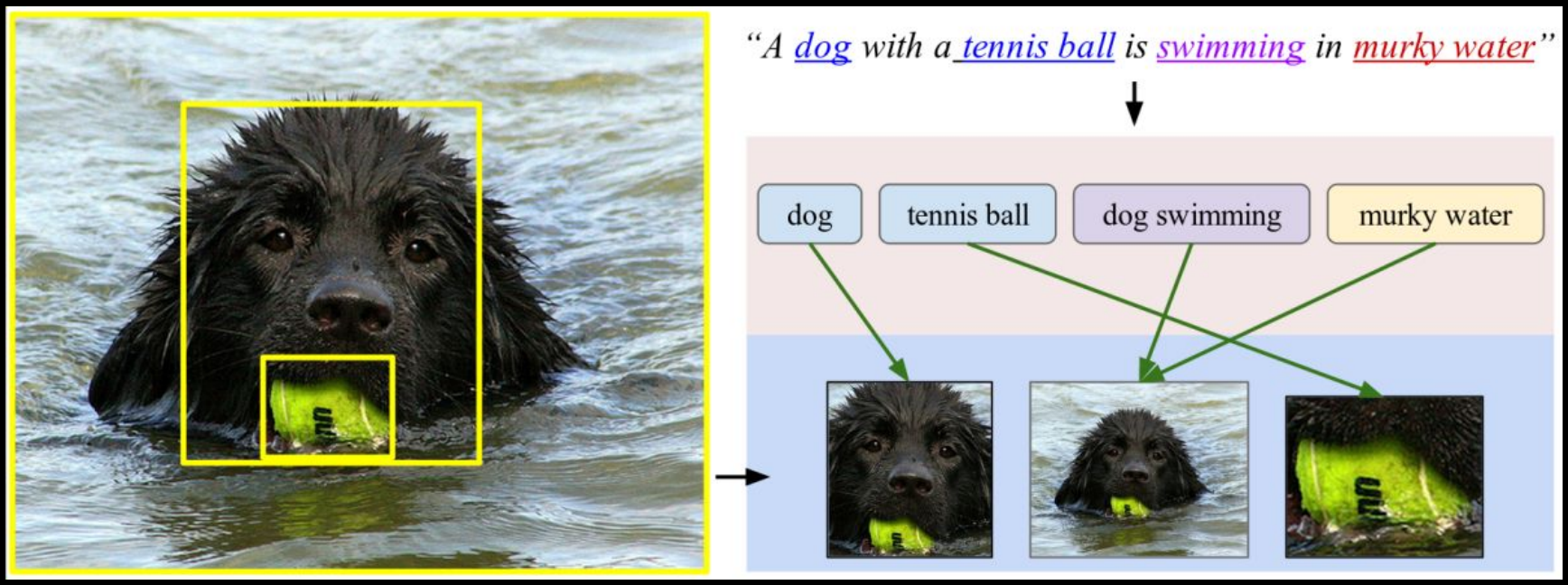


Mapeamento Bidirecional Imagem-Sentença

KARPATHY, Andrej; JOULIN, Armand; LI, Fei Fei F. Deep fragment embeddings for bidirectional image sentence mapping. In: **Advances in neural information processing systems**. 2014. p. 1889-1897.

Mapeamento Bidirecional Imagem-Sentença

Modelo que pega um conjunto de imagens e suas descrições e aprende a associar seus fragmentos.



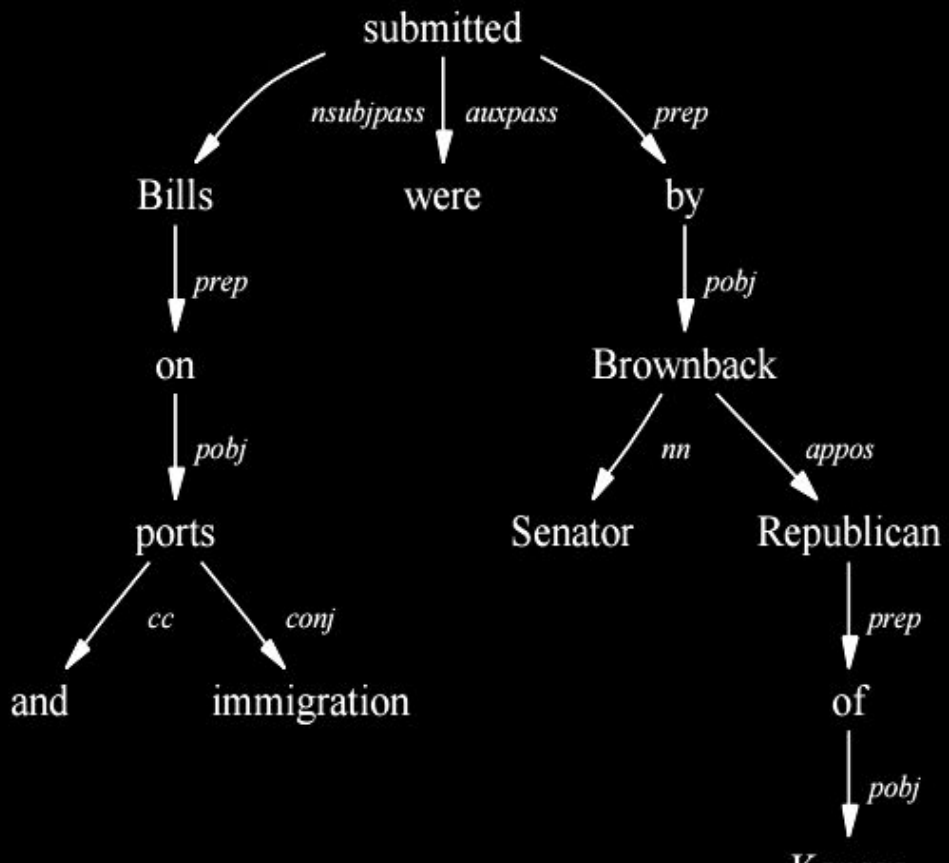
Mapeamento Bidirecional Imagem-Sentença

- Em uma imagem, um fragmento corresponde a detecção de objetos e contexto da cena.

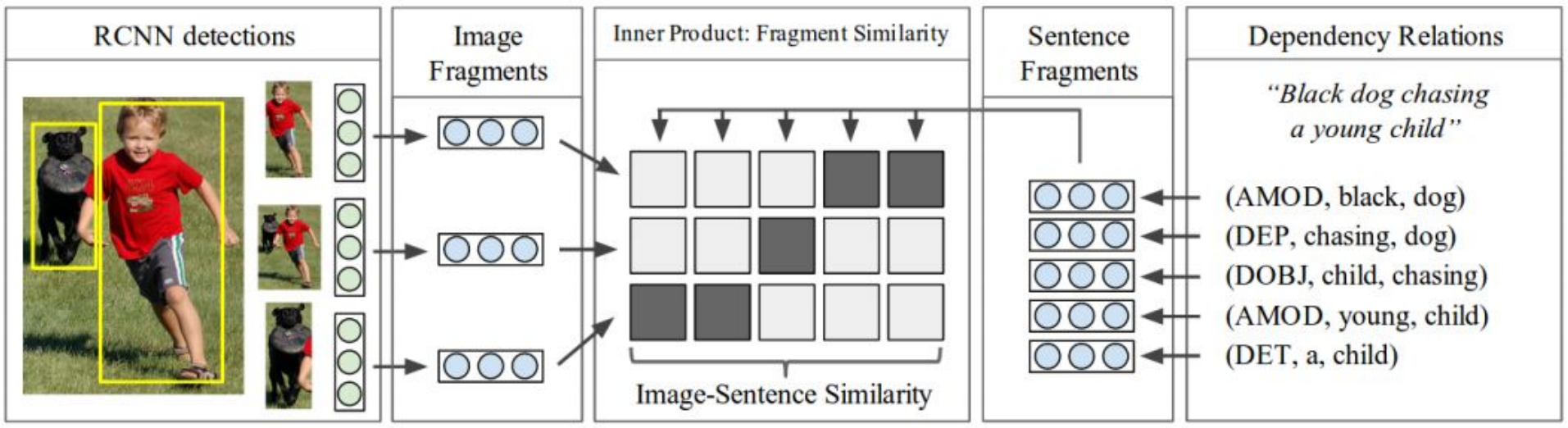
- Em sentenças, fragmentos consistem em relações em uma árvore de dependências tipada (*typed dependency tree*).

Árvore de Dependências Tipada

"Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."



Mapeamento Bidirecional Imagem-Sentença



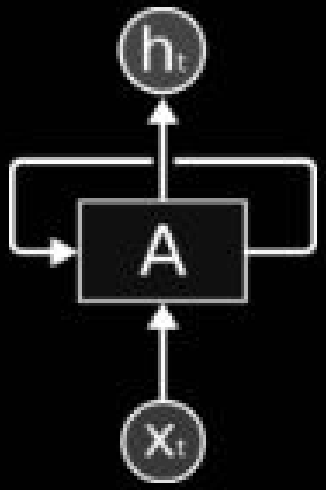
Mapeamento Bidirecional Imagem-Sentença

Mapeamento Bidirecional Imagem-Sentença

40.7 (DOBJ, sunglasses, wearing)
20.6 (DET, a, baby)
19.0 (NSUBJ, baby, sits)
18.6 (PREP ON, lap, sits)
10.6 (VMOD, wearing, baby)
8.1 (AMOD, small, baby)
7.9 (POSS, adult, lap)
6.0 (DET, an, adult)

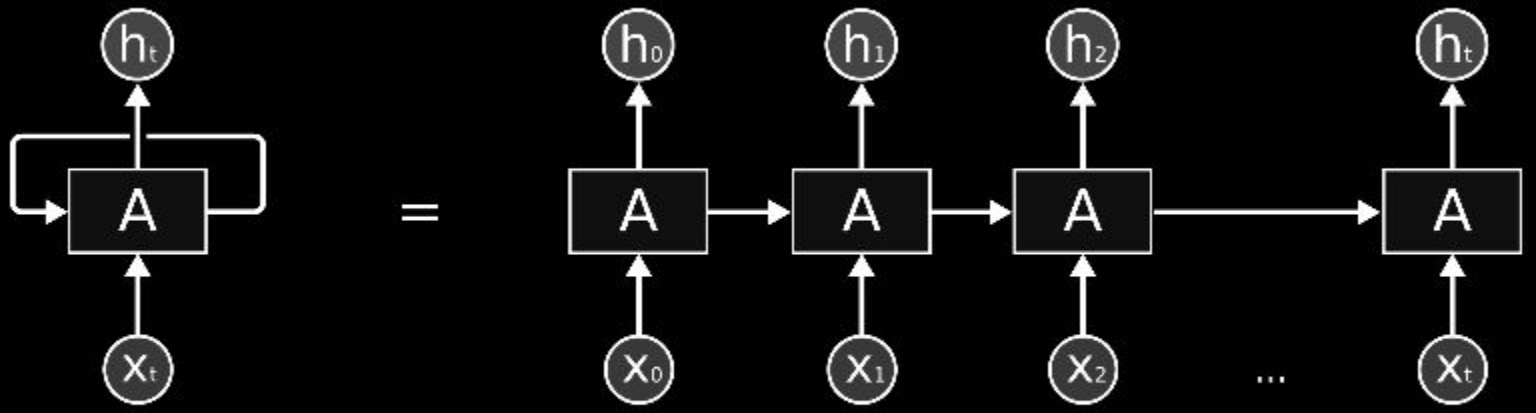
1. A small baby wearing sunglasses sits on an adult's lap
2. A woman holds a fat baby with sunglasses and a hat
3. A naked toddler is covering a naked baby with paint
4. A naked baby and toddler on the floor covered in paint, the toddler putting her hands on the baby 's head
5. A woman is holding onto a baby wearing who is wearing sunglasses

Redes Neurais Recorrentes (RNN)



Recurrent Neural Networks have loops.

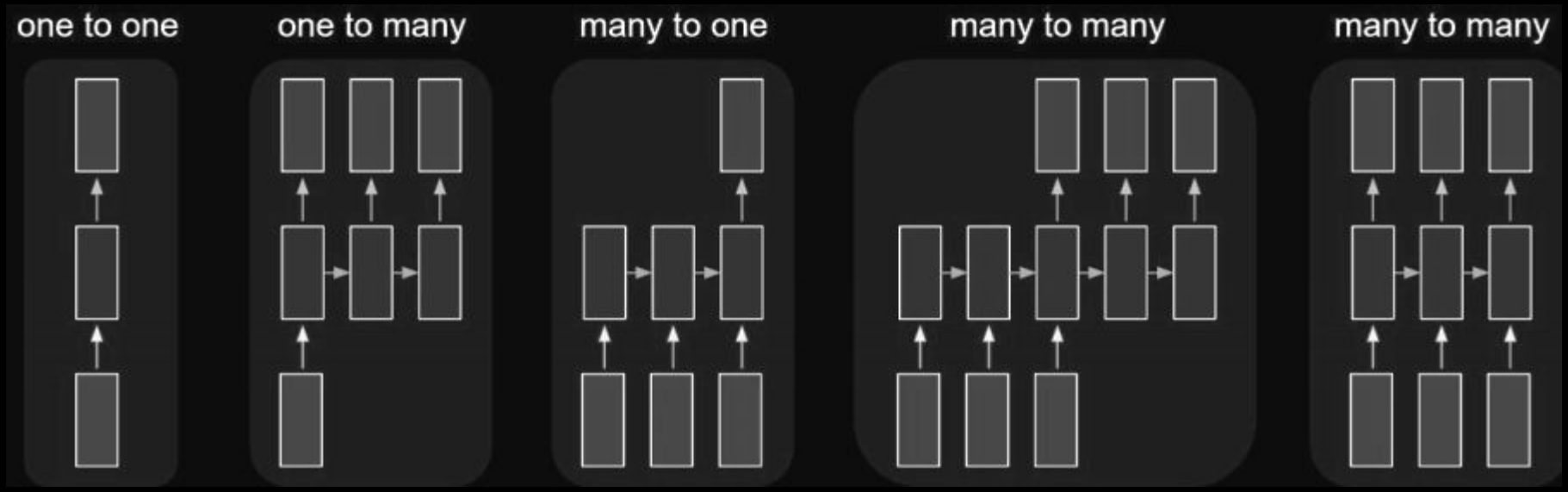
Redes Neurais Recorrentes (RNN)



An unrolled recurrent neural network.

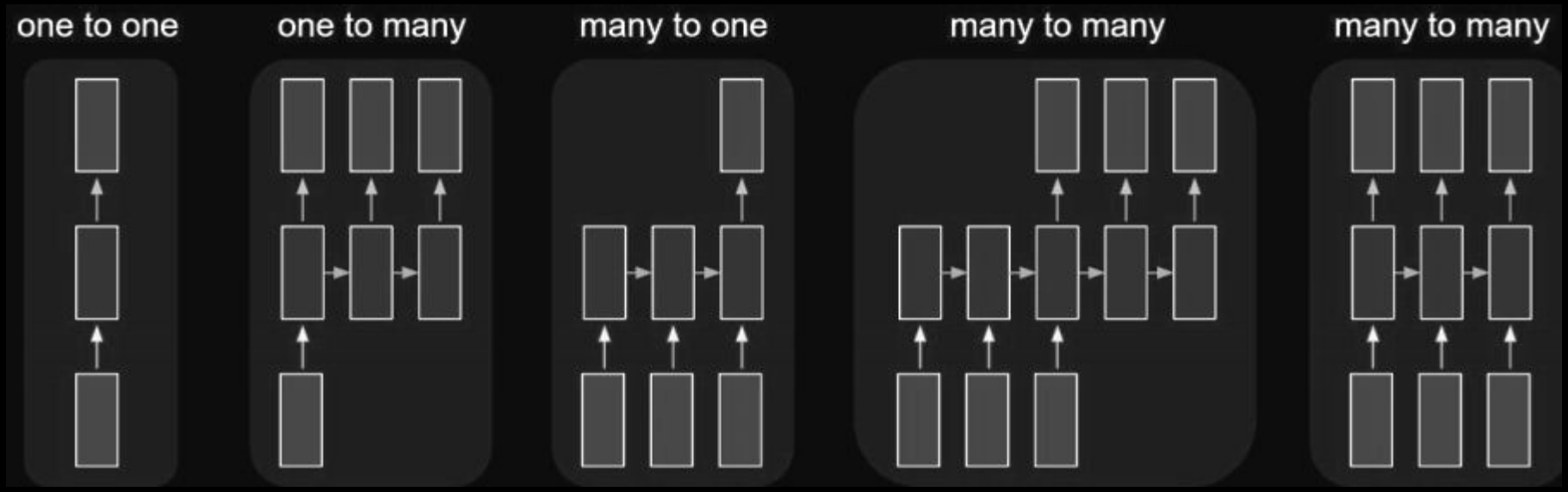
Redes Neurais Recorrentes (RNN)

As RNN são flexíveis e atendem a várias aplicações:



Redes Neurais Recorrentes (RNN)

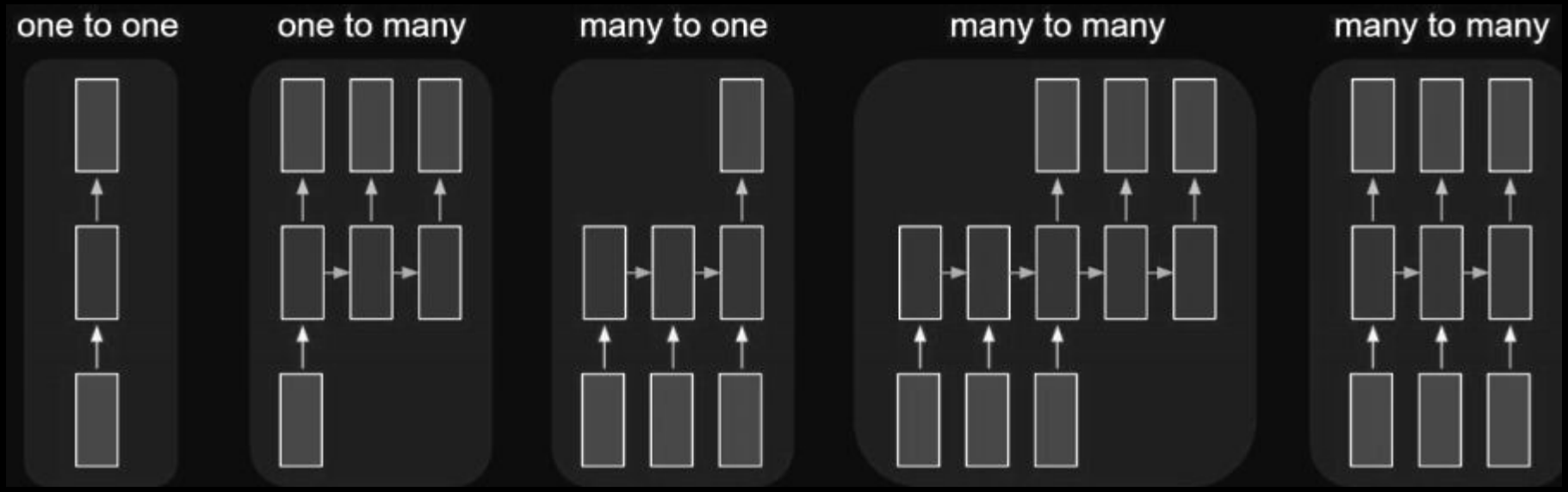
As RNN são flexíveis e atendem a várias aplicações:



Previsão de palavras para formulação de frases

Redes Neurais Recorrentes (RNN)

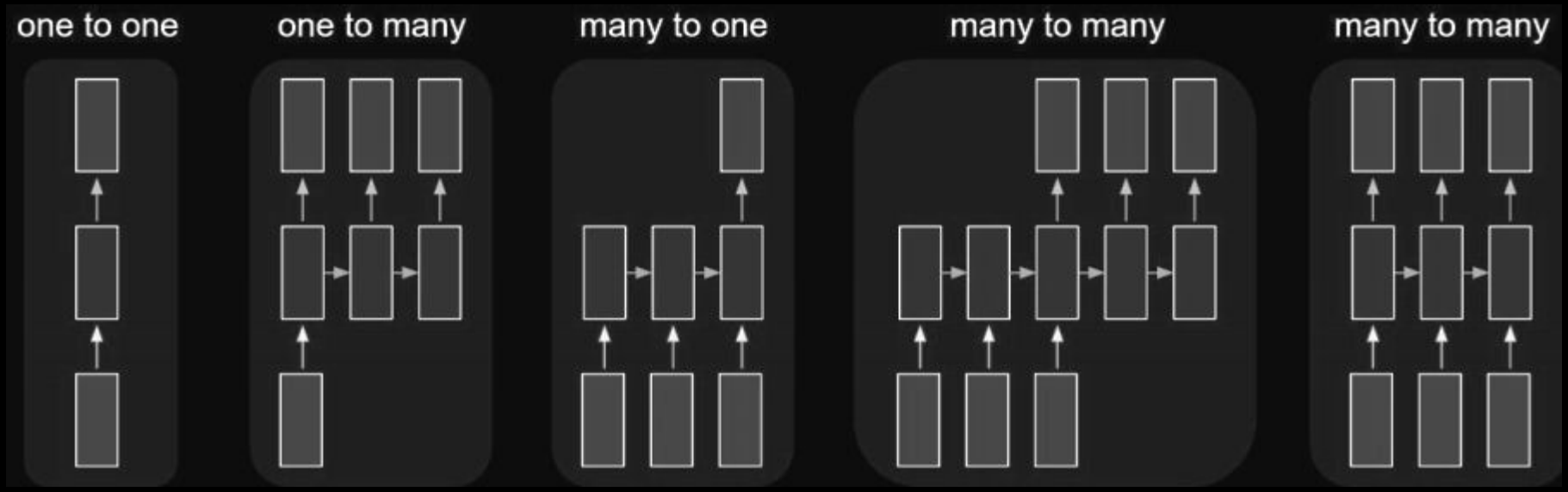
As RNN são flexíveis e atendem a várias aplicações:



Construção de Legendas para Imagens

Redes Neurais Recorrentes (RNN)

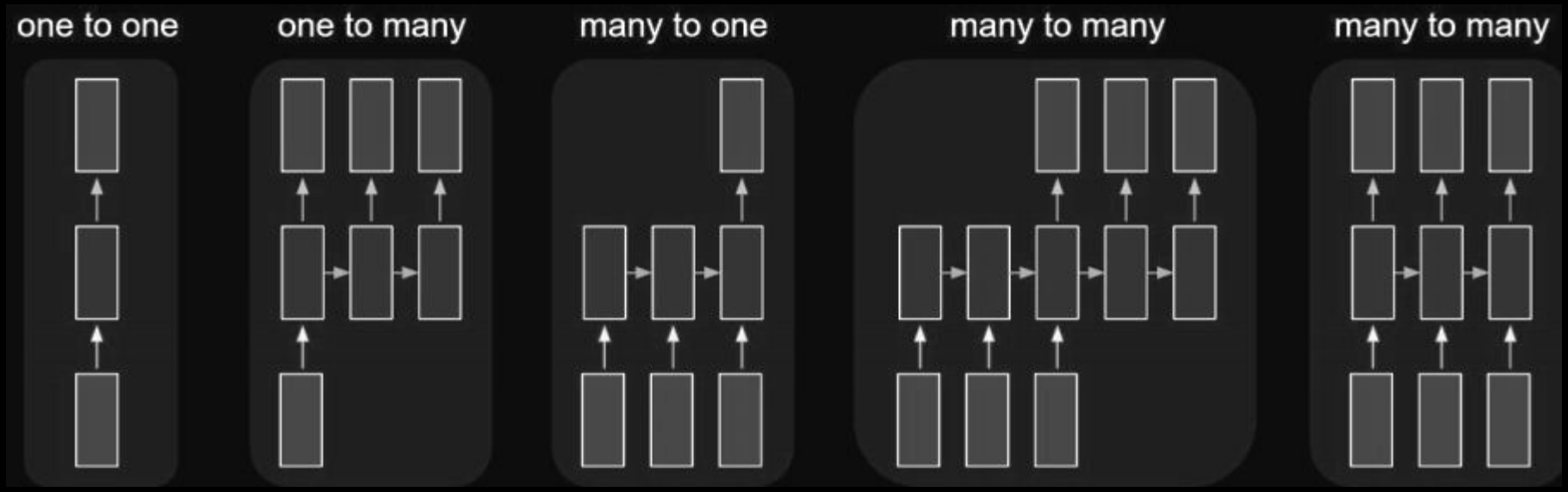
As RNN são flexíveis e atendem a várias aplicações:



Classificação de Sentimentos.

Redes Neurais Recorrentes (RNN)

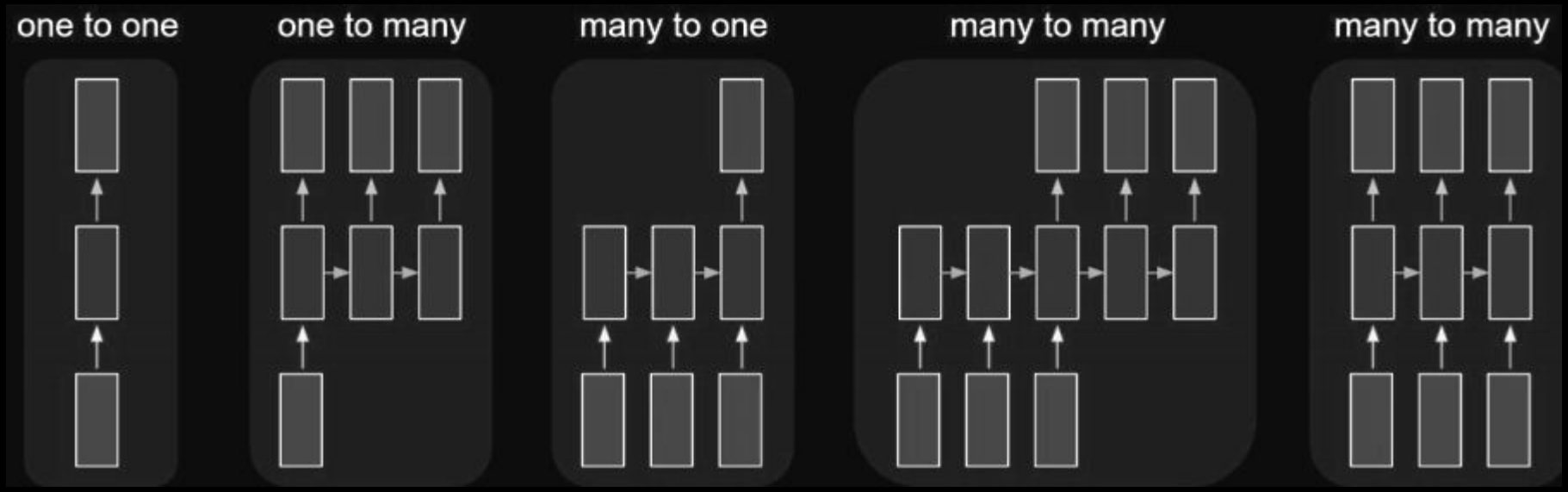
As RNN são flexíveis e atendem a várias aplicações:



Tradução Automática

Redes Neurais Recorrentes (RNN)

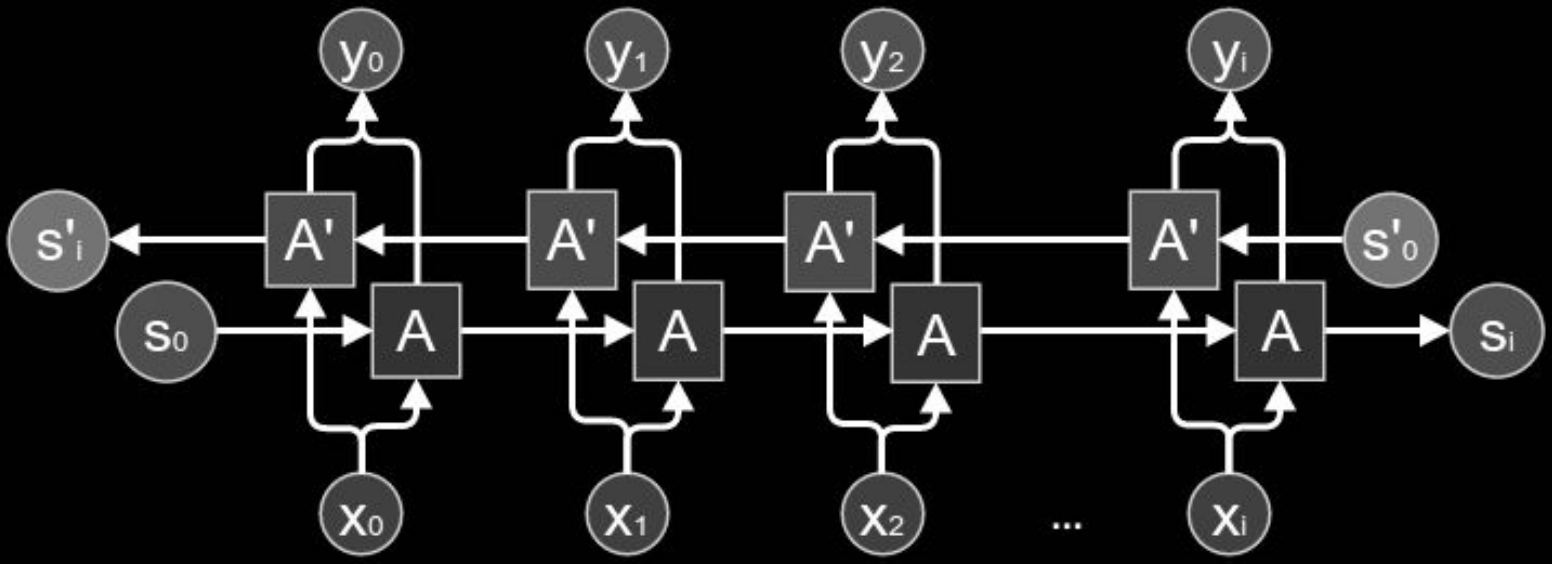
As RNN são flexíveis e atendem a várias aplicações:



Classificação de vídeos no nível de quadros

Redes Neurais Recorrentes Bidirecionais (BRNN)

As RNN são flexíveis e atendem a várias aplicações:



Gerando descrição de imagens com R-CNN e BRNN

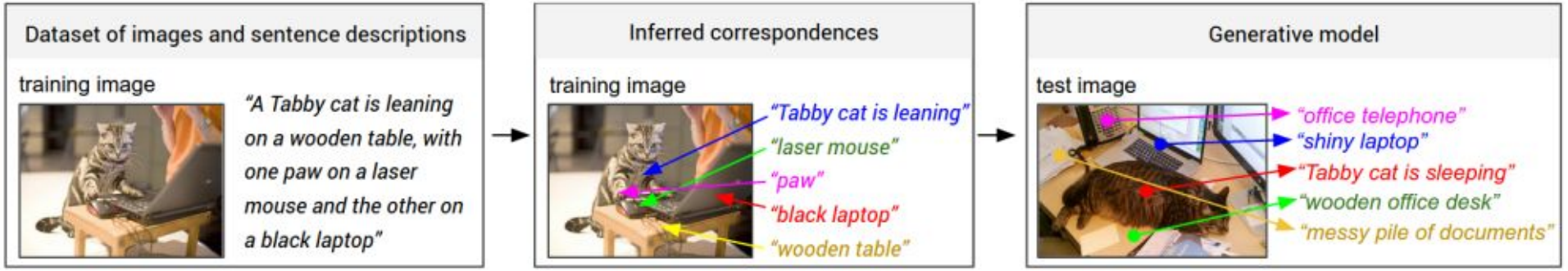
KARPATHY, Andrej; FEI-FEI, Li. Deep visual-semantic alignments for generating image descriptions. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2015. p. 3128-3137.

Gerando descrição de imagens com R-CNN e BRNN

- Mesma R-CNN usada no trabalho anterior.
- A proposta é substituir a árvore de dependências por uma BRNN.
- A BRNN recebe como “viés” a saída da última camada completamente conectada da R-CNN (fully connected).

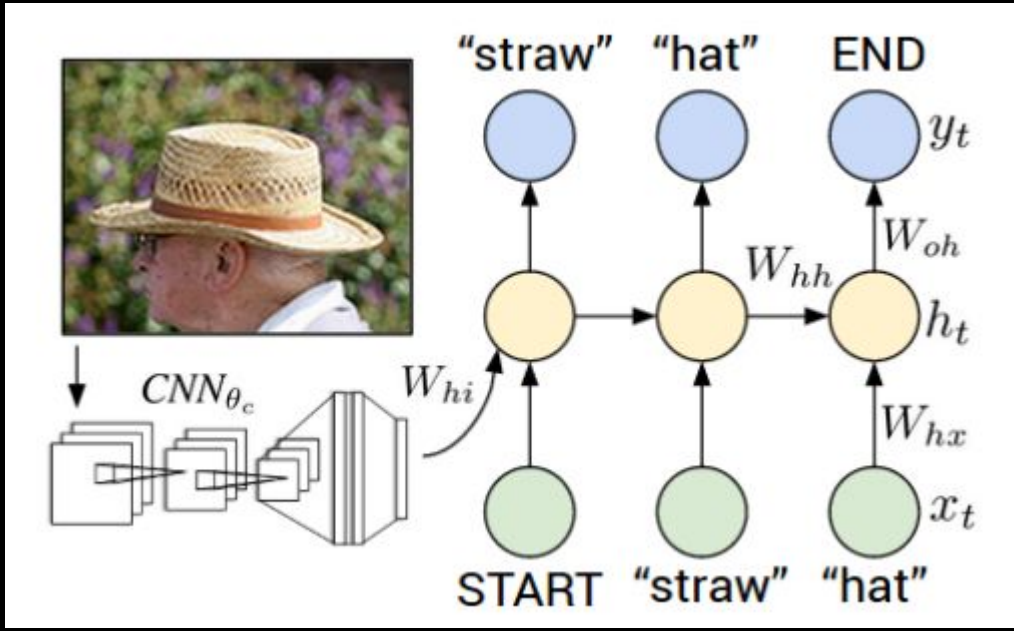
Gerando descrição de imagens com R-CNN e BRNN

- Abordagem:
  Treinamento: Entrar com imagens e suas descrições.
  Treinamento: Inferir correspondências.
  Teste: gerar novas descrições.



Gerando descrição de imagens com R-CNN e BRNN

No teste a saída da CNN e um token inicial (START) são dados. A BRNN prediz as próximas palavras até um token final (END).



Gerando descrição de imagens com R-CNN e BRNN.

Imagem de teste

Gerando descrição de imagens com R-CNN e BRNN.

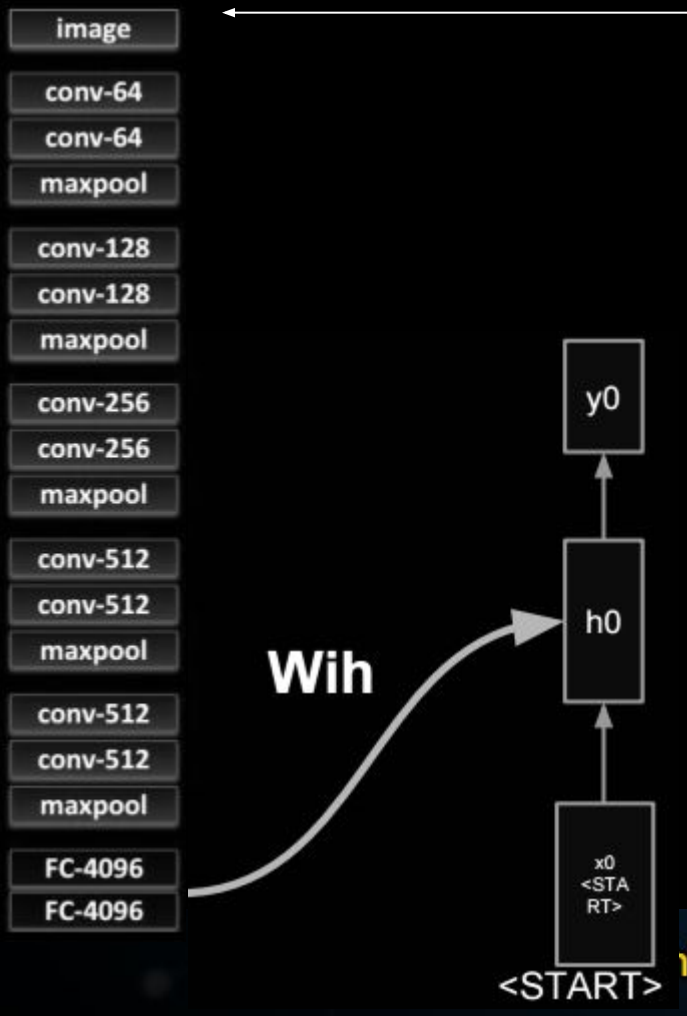
Imagem de teste

Gerando descrição de imagens com R-CNN e BRNN.

Imagem de teste

Gerando descrição de imagens com R-CNN e BRNN.

Imagem de teste

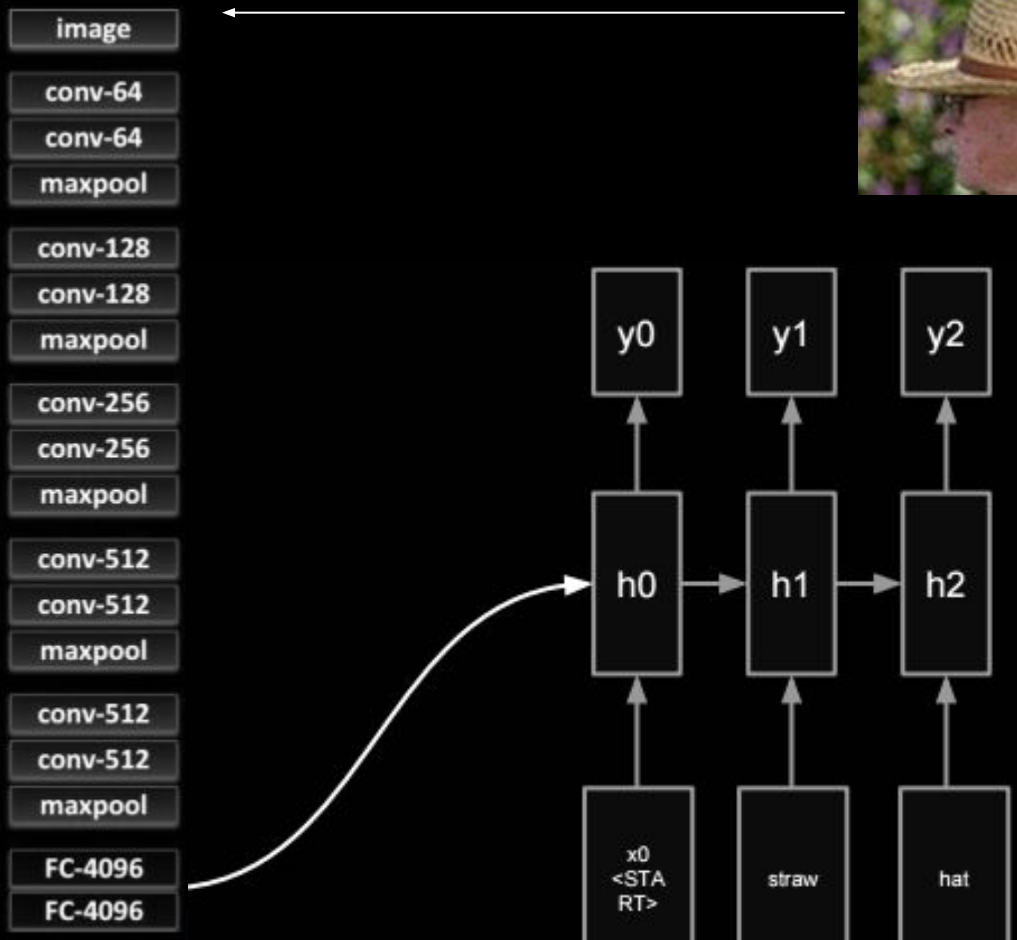


$$h = \tanh(W_{xh} * x + W_{hh} * h)$$

$$h = \tanh(W_{xh} * x + W_{hh} * h + W_{ih} * v)$$

Gerando descrição de imagens com R-CNN e BRNN.

Imagem de teste



Gerando descrição de imagens com R-CNN e BRNN.

man in black shirt is playing guitar.

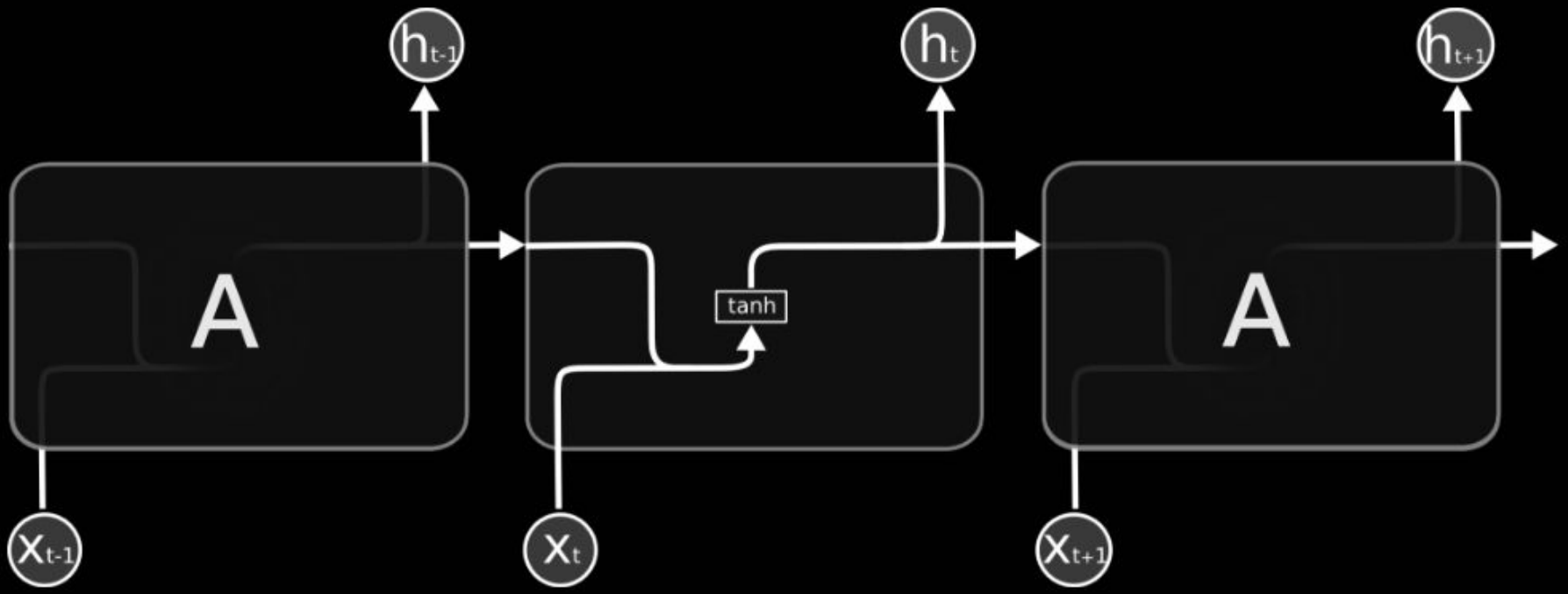
construction worker in orange safety vest is working on road.

two young girls are playing with lego toy.

Gerando descrição de imagens com R-CNN e BRNN.

Exemplos (predição de regiões):

Redes Neurais Recorrentes (RNN)



The repeating module in a standard RNN contains a single layer.

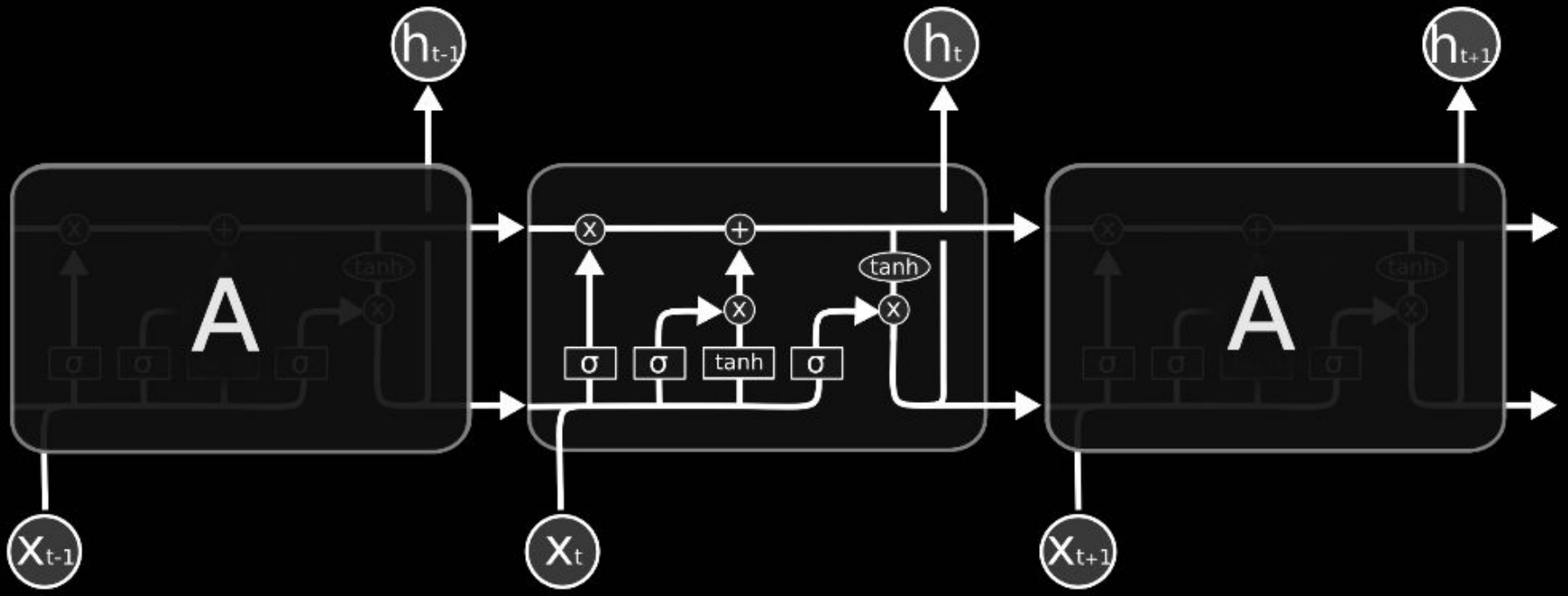
Redes Neurais Recorrentes (RNN)

- Assim como as redes neurais comuns, as RNN sofrem o vanishing gradient.

- A predição abaixo requer pouca “memória”:
  “As nuvens estão no céu”.

- Todavia, a predição abaixo é muito mais complexa:
  “Eu cresci no Brasil... Eu falo fluentemente Português.”

Long Short-Term Memory (LSTM)



The repeating module in an LSTM contains four interacting layers.

Long Short-Term Memory (LSTM)

- Cada camada adicionada é chamada de gate.

  Forget gate (f): controla se o valor da célula atual é para ser esquecido.
  Input gate (i): controla se o valor da entrada atual é para ser lido.
  Output gate (o): controla se é para retornar o novo valor de saída.

Gerando Legendas de Imagens com LSTM

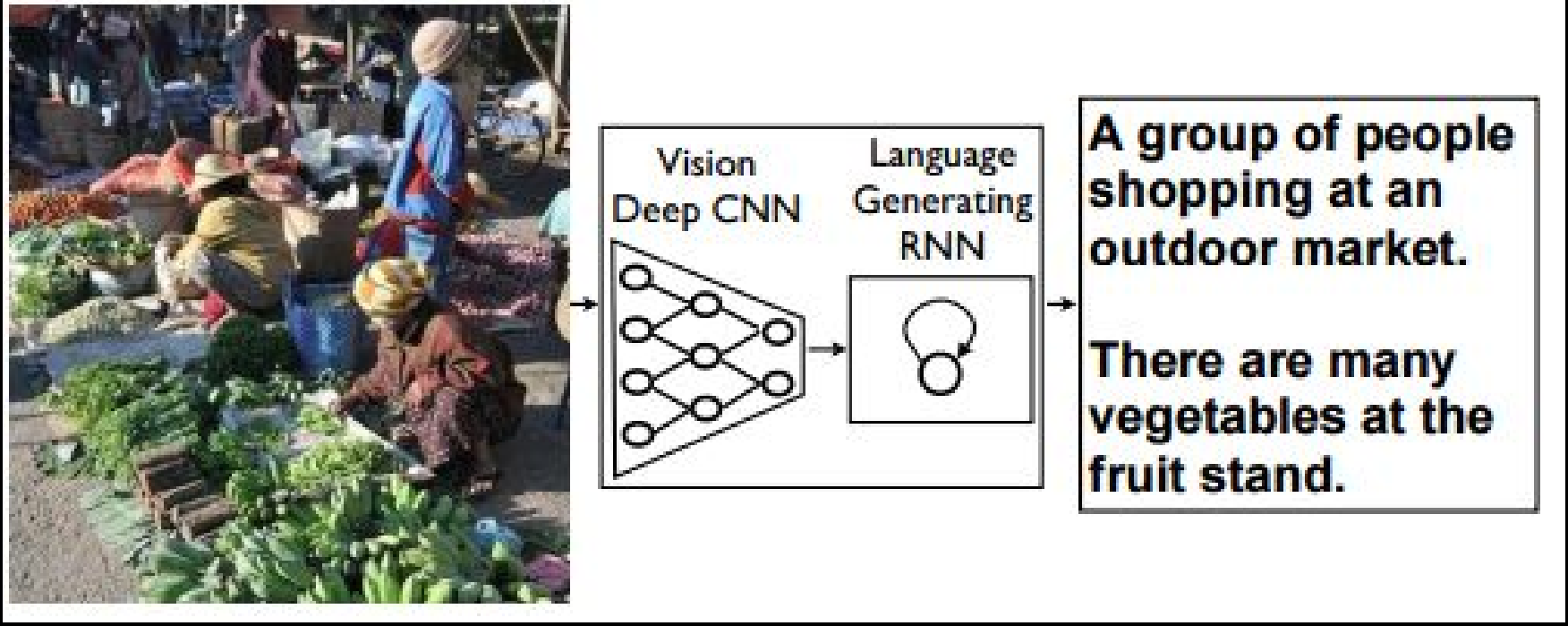
VINYALS, Oriol et al. Show and tell: A neural image caption generator. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2015. p. 3156-3164.

VINYALS, Oriol et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2016.

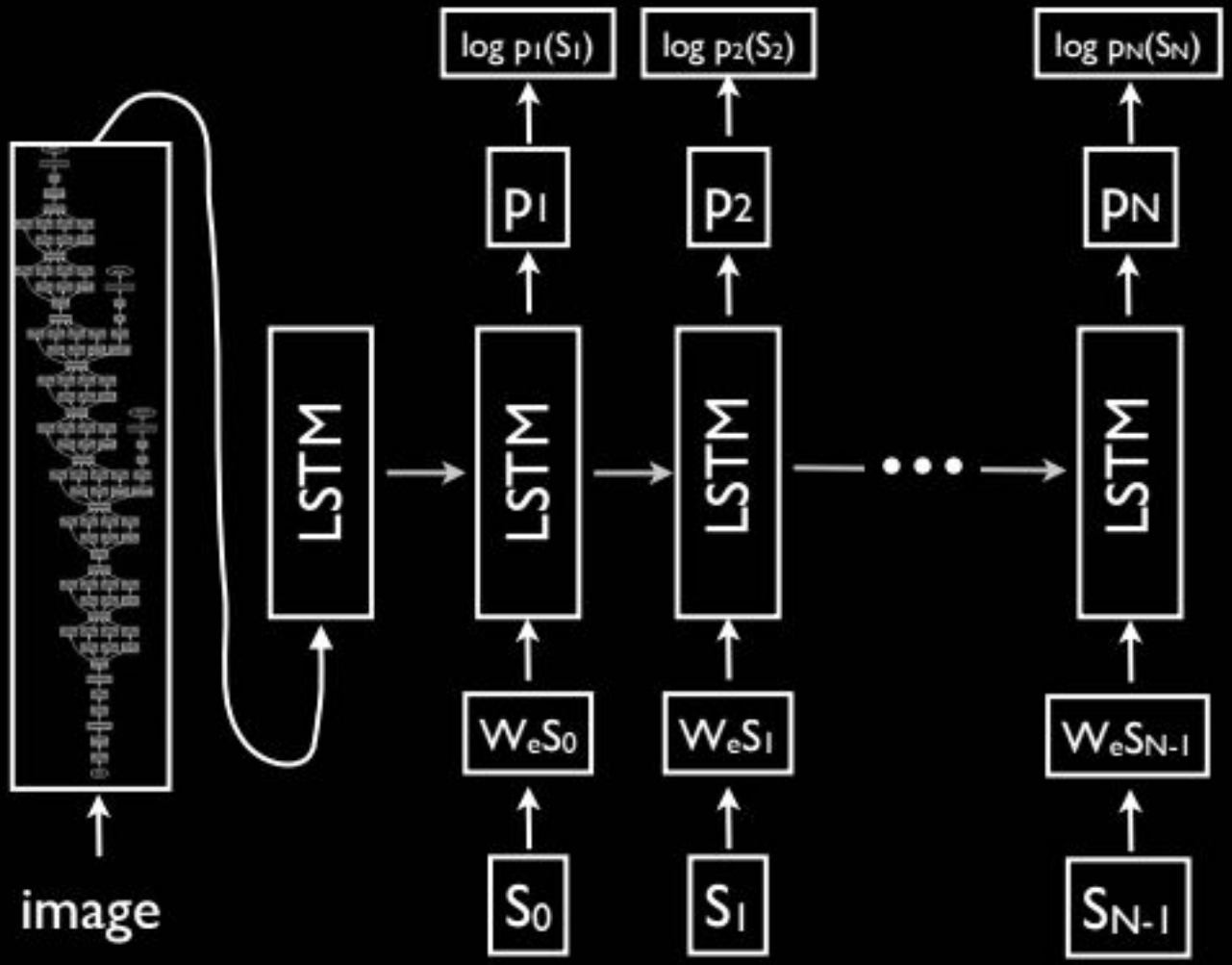
Gerando Legendas de Imagens com LSTM

- Proposta: utilizar o mesmo mecanismo utilizado na tradução de textos.
- A entrada, ao invés de ser um texto em uma língua, é a imagem (saída da CNN).
- Não utiliza R-CNN como as outras propostas.

Gerando Legendas de Imagens com LSTM



Gerando Legendas de Imagens com LSTM



DEEP LEARNING BRASIL
SUMMER SCHOOL

INSTITUTO DE INFORMÁTICA UFG
FASAM FACULDADE SUL-AMERICANA

Patrocínio: ZG SOLUÇÕES, Saúdemobi, INFOMACH TECNOLOGIA PARA NEGÓCIOS

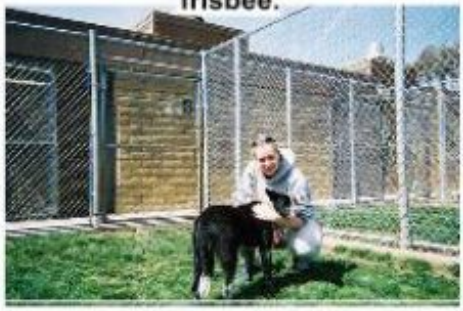
Apoio: NVIDIA, DATA H

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

Mecanismo de Atenção para Legendar Imagens

LU, J.; XIONG , C.; PARIKH, D.; SOCHER, R. **Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning**. IEEE Conference on Computer Vision and Pattern Recognition Knowing. 2017.

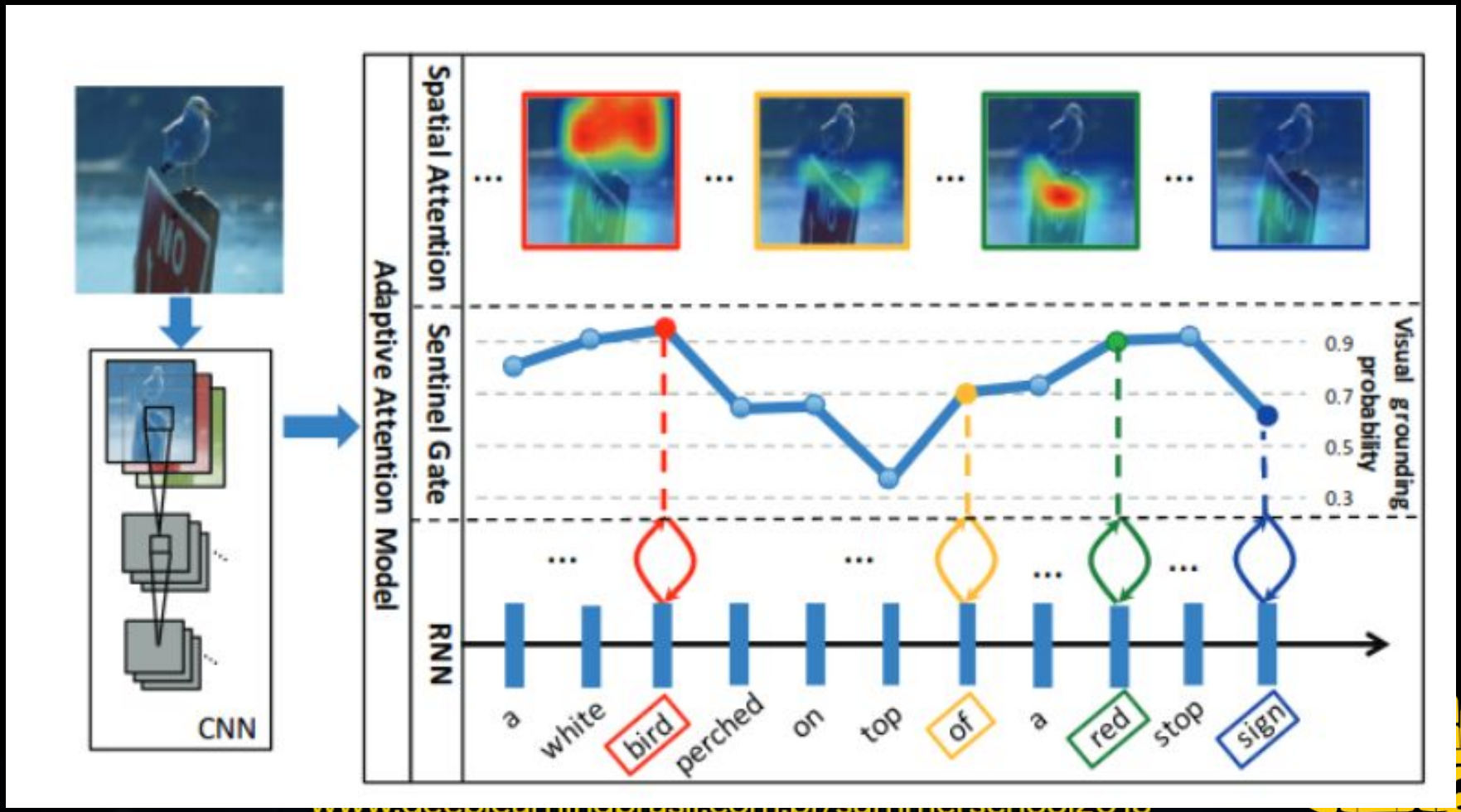
FU, K.; JIN, J.; CUI, R.; SHA, F.; ZHANG , C. **Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts**. IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1–1, 2017.

Atenção Adaptativa com LSTM

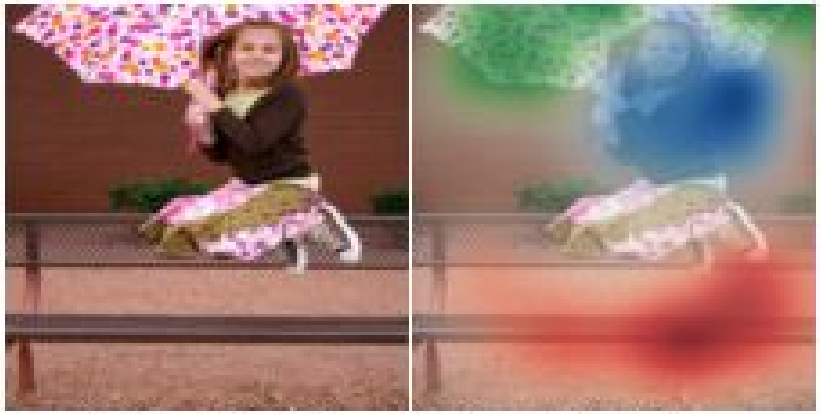
- Novo vetor interno ao LSTM: "sentinela visual".

- Novo *Sentinel Gate*: controla se para dar maior ou menor atenção à imagem (através do vetor sentinela visual).

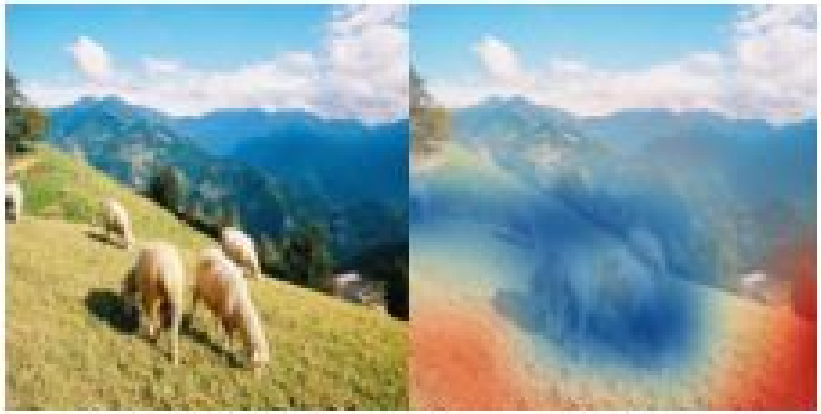
Atenção Adaptativa com LSTM



Atenção Adaptativa com LSTM



a little girl sitting on a bench holding an umbrella.



a herd of sheep grazing on a lush green hillside.

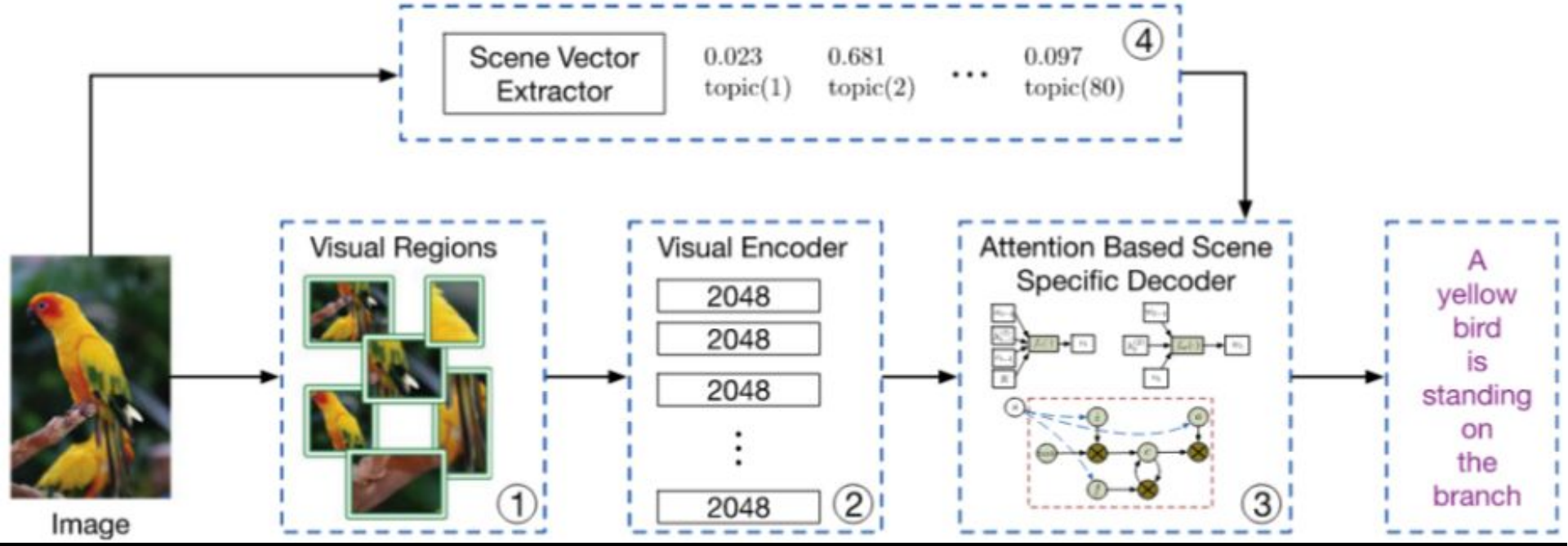
a yellow plate topped with meat and broccoli.

a zebra standing next to a zebra in a dirt field.

Atenção baseada em região e contextos específicos de cena.

- Múltiplas regiões visuais das quais as características visuais são extraídas (não é R-CNN).
- LSTM alimentada pelas características visuais.
- Prediz:
  - sequência de focus em diferentes regiões.
  - sequência de palavras de contexto específico.

Atenção baseada em região e contextos específicos de cena.



Atenção baseada em região e contextos específicos de cena.

[RA+SS]
a man riding a wave on top of a surfboard

human:
1) a man riding a board on top of a wave in the ocean
2) a guy in a black and white outfit is surfing
3) a man in a wet suit riding a surfboard on a wave
4) a male surfing a large ocean wave on a white surfboard
5) a man is riding a surboard on a wave

[RA+SS]
a tall tower with a clock on it

human:
1) a tall tower with a clock stands above a winter sky
2) there is a tree next to the clock tower
3) a large clock tower on a cloudy winter day
4) there is a clock in the center of a tower
5) a tower that will have a large clock at the top

Referências

VINYALS, Oriol et al. Show and tell: A neural image caption generator. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2015. p. 3156-3164.

VINYALS, Oriol et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2016.

KARPATHY, Andrej; FEI-FEI, Li. Deep visual-semantic alignments for generating image descriptions. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2015. p. 3128-3137.

KARPATHY, Andrej; JOULIN, Armand; LI, Fei Fei F. Deep fragment embeddings for bidirectional image sentence mapping. In: **Advances in neural information processing systems**. 2014. p. 1889-1897.

Referências

LU, J.; XIONG , C.; PARIKH, D.; SOCHER, R. **Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning**. IEEE Conference on Computer Vision and Pattern Recognition Knowing. 2017.

FU, K.; JIN, J.; CUI, R.; SHA, F.; ZHANG , C. **Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts**. IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1–1, 2017.

CHEN, L.; ZHANG, H.; XIAO, J.; NIE, L.; SHAO, J.; LIU, W.; CHUA, T.-S. **SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning**. 2017

Referências

DE MARNEFFE, Marie-Catherine et al. Generating typed dependency parses from phrase structure parses. In: **Proceedings of LREC**. 2006. p. 449-454.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, v. 9, n. 8, p. 1735-1780, 1997.

OLAH, Christopher. Understanding LSTM Networks. **Net: http://colah. github. io/posts/2015-08-Understanding-LSTMs**, 2015.