Sheffield MultiMT: Using Object Posterior Predictions for Multimodal Machine Translation

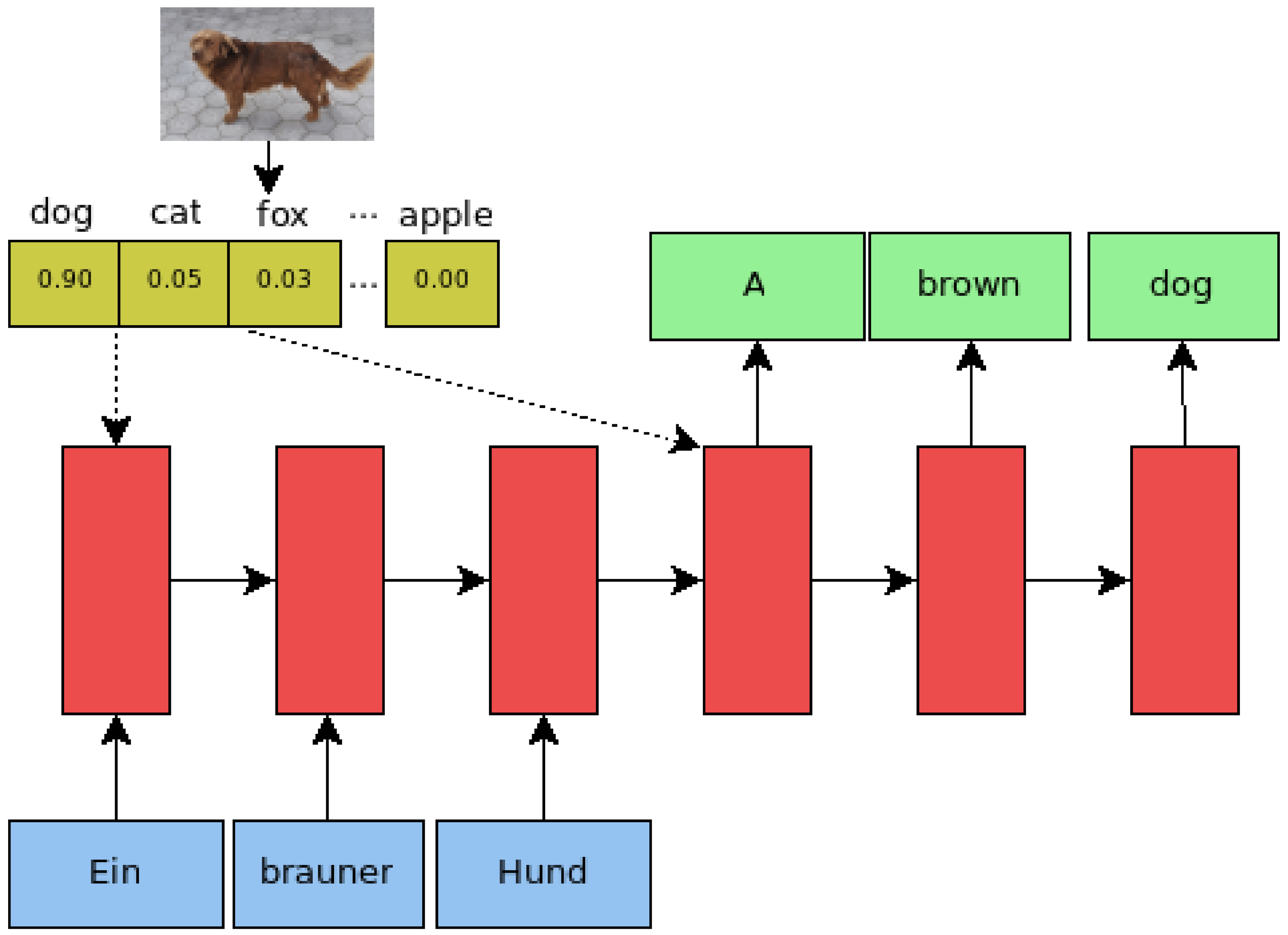
The University Of Sheffield.
multimt

Pranava Madhyastha
Josiah Wang
Lucia Specia

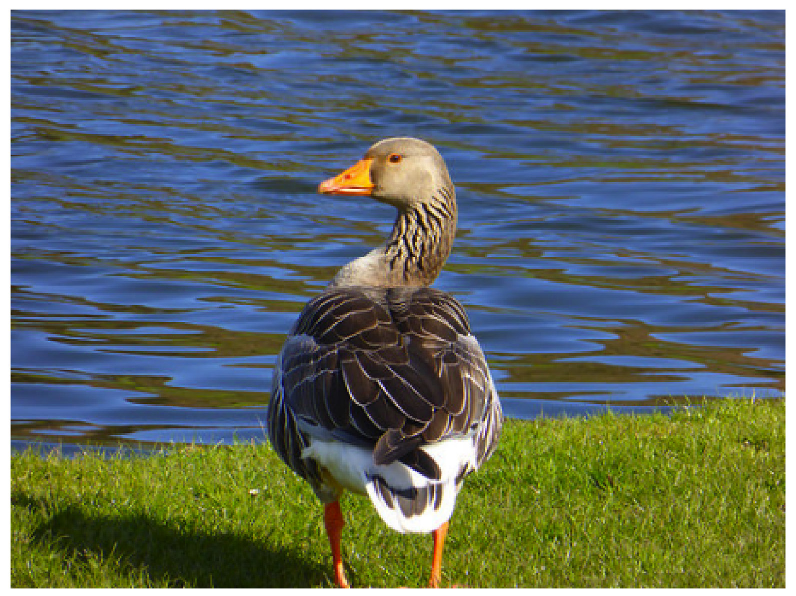
— In Focus —

- Question: Can higher level semantic attributes play a role in multimodal MT?
- Proposal:
  1. Exploit class predictions directly from pre-trained SOTA image network
  2. Class predictions are over 1,000 WordNet synsets (ILSVRC image classification challenge)
  3. Evaluation with low-level and high-level features

— A Quick Look at the System —



— Glimpse of the Results —



EN	a duck on the bank of a river
DE (**Baseline**)	eine ente an der küste eines flusses .
DE (**InitDec**)	eine ente am ufer eines flusses
DE (Reference)	eine ente am ufer eines flusses
FR (**Baseline**)	un canard sur l' eau , dans une rivière
FR (**InitDec**)	un canard sur la rive d' une rivière
FR (Reference)	un canard sur la berge d' une rivière

- Top-5 Class Predictions: (i) goose, (ii) drake, (iii) european gallinule, (iv) merganser, (v) black swan

— System Description —

- Visual Features: Posterior class distribution from ResNet512
- Standard LSTM attention-based encoder decoder architecture
- Greedy decoding
- Constrained setting

— Conditioning Image Info —

Class Predictions: Using *Softmax*

- InitEnc: Image features initializing the *encoder*
- InitDec: Image features initializing the *decoder*
- Proj: Projected image features added to *each input token on the encoder*

Image features projected to smaller dimensionality followed by a ReLU non linearity

— Setup and Hyperparametres —

Image Features:	1000-dim posterior distribution
Model:	Embedding: 128-dim; Hidden: 256-dim; Adadelta optimizer
Word Threshold:	Words appearing at least twice
Training:	Batch of 20; 50 epochs
Comparision:	Baseline: (i) Standard NMT, (ii) Pool5 features as image info
<UNK> handling:	Replace with an empty string

Important: the posteriors may contain distributions for *non-caption relevant categories*

— Results: Flickr test data —

	Flickr Feature	Model	Meteor	BLEU
EN–DE	–	Text-only	43.7	24.4
	Pool5	Proj	–	–
		InitEnc	43.0	23.5
		InitDec	44.3	24.6
	Softmax	Proj	43.4	24.2
		InitEnc	42.4	23.3
		InitDec	44.5	25.0
EN–FR	–	Text-only	62.2	44.2
	Pool5	Proj	–	–
		InitEnc	61.1	43.5
		InitDec	61.0	43.4
	Softmax	Proj	61.5	43.6
		InitEnc	61.0	43.3
		InitDec	62.8	45.0

— Results: MSCOCO test data —

	MSCOCO Feature	Model	Meteor	BLEU
EN–DE	–	Text-only	39.6	20.7
	Pool5	Proj	–	–
		InitEnc	39.1	20.4
		InitDec	39.5	20.4
	Softmax	Proj	40.0	21.0
		InitEnc	37.5	18.8
		InitDec	40.7	21.4
EN–FR	–	Text-only	57.4	37.2
	Pool5	Proj	–	–
		InitEnc	56.7	36.5
		InitDec	56.7	36.9
	Softmax	Proj	57.0	36.8
		InitEnc	55.5	35.5
		InitDec	57.3	37.2

— Observations —

- Class predictions seem to outperform Pool5 features
- On average, superior performance is observed when image features are used to condition the decoder
- Manual inspection suggests the class predictions correlate with the generated outputs
- Fine-tuning the class predictions and tuning system components should lead to improvements
- Not directly comparable to most other submissions

Acknowledgements: This work was supported by the MultiMT project (EU H2020 ERC Starting Grant No. 678017)