Multimodal Machine Translation

鲍建竹

- Multimodal Attention for Neural Machine Translation (COLING 2016)
- Probing the Need for Visual Context in Multimodal Machine Translation (NAACL 2019)

  - Motivation
  - Method
  - Experiment
  - Pros and cons
  - Inspiration

Multimodal Attention for Neural Machine Translation

Ozan Caglayan[1,2], **Loïc Barrault**[1], **Fethi Bougares**[1]
[1] LIUM, University of Le Mans / France
[2] Galatasaray University / Turkey
[1]FirstName.LastName@univ-lemans.fr
[2]ocaglayan@gsu.edu.tr

1. Motivation

- Why?
  - Dealing with multimodal stimuli in order to perceive the surrounding environment and to understand the world is natural for human beings.
  - It is not the case for artificial intelligence.

- Current work
  - Neural machine translation (NMT)
  - Multilingual information
  - Image captioning

- an NMT enriched with convolutional image features as auxiliary source representation
- or an image captioning system producing image descriptions in a language T, supported with source descriptions in another language S.

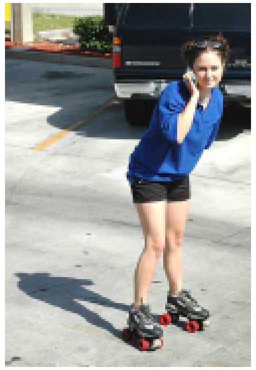
2. Method

2.1 The Multi30K Dataset

1. Brick layers constructing a wall.
2. Maurer bauen eine Wand.

1. The two men on the scaffolding are helping to build a red brick wall.
2. Zwei Mauerer mauern ein Haus zusammen.

1. Trendy girl talking on her cellphone while gliding slowly down the street
2. Ein schickes Mädchen spricht mit dem Handy während sie langsam die Straße entlangschwebt.



1. There is a young girl on her cell-phone while skating.
2. Eine Frau im blauen Shirt telefoniert beim Rollschuhfahren.

(a) Translations

(b) Independent descriptions

- 31K images
- With 5 English descriptions
- With 5 independently descriptions in German

2.2 Multimodal Machine Translation (MMT)

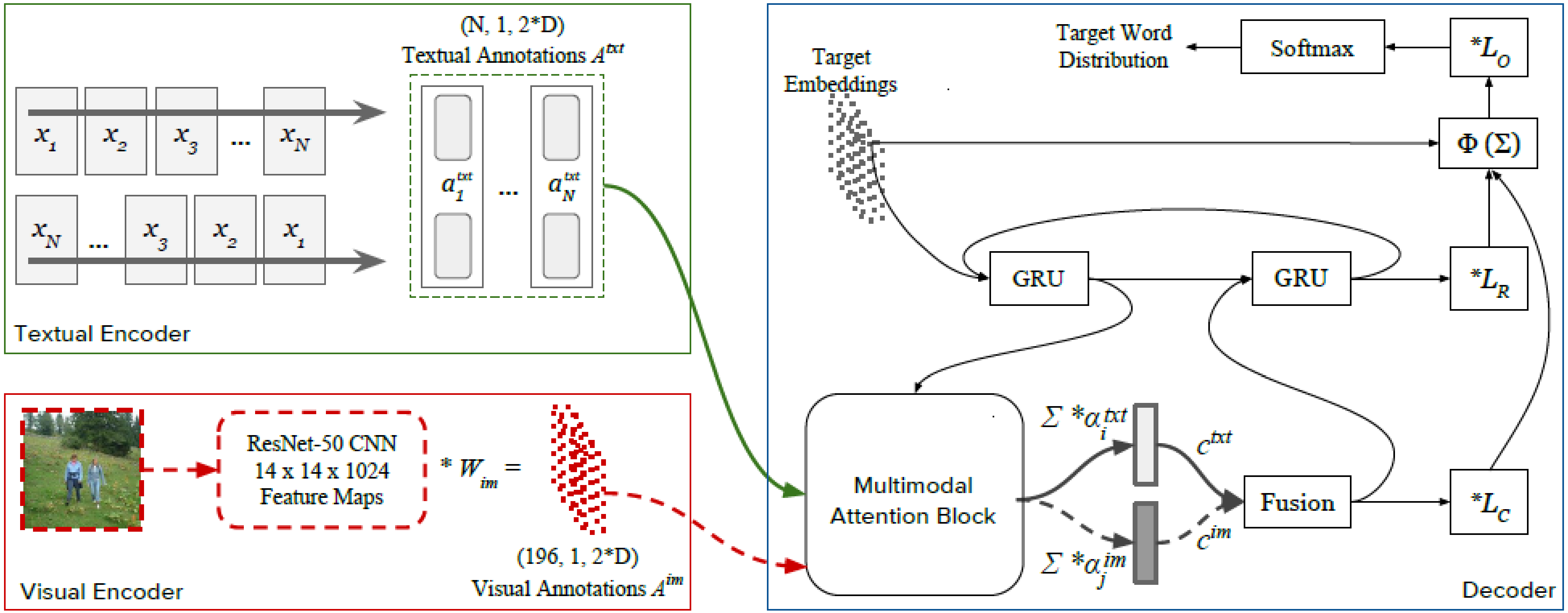
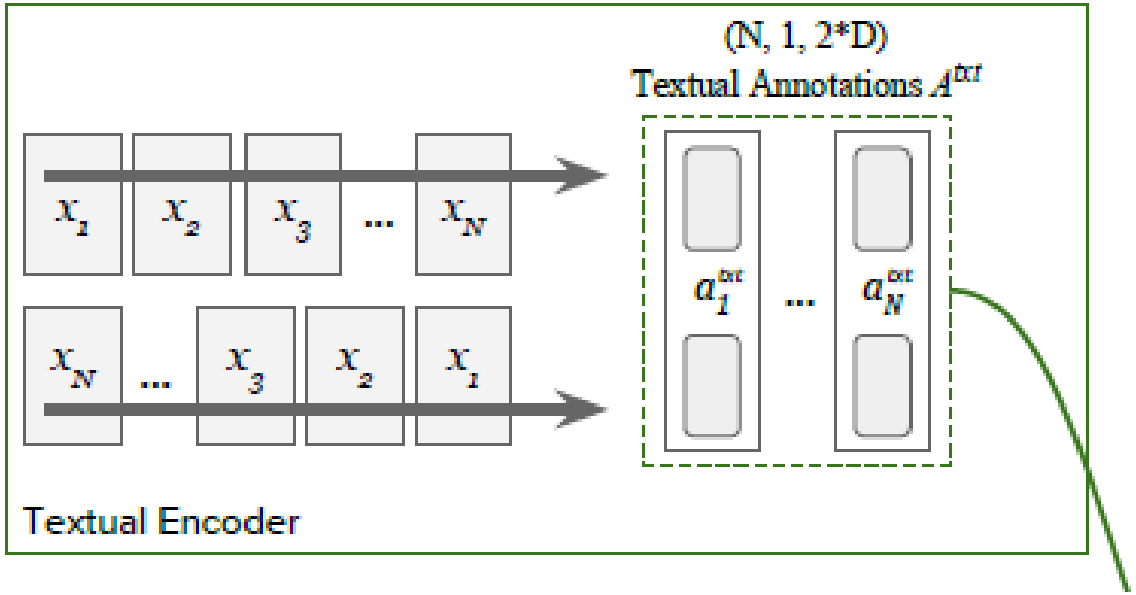


Figure 1: The architecture of MNMT: The boxes with $*$ refer to a linear transformation while $\Phi(\Sigma)$ means a $tanh$ applied over the sum of the inputs.

2.2.1 Textual Encoder



$$X = (x_1, x_2, ..., x_N), x_i \in \mathbb{R}^E$$

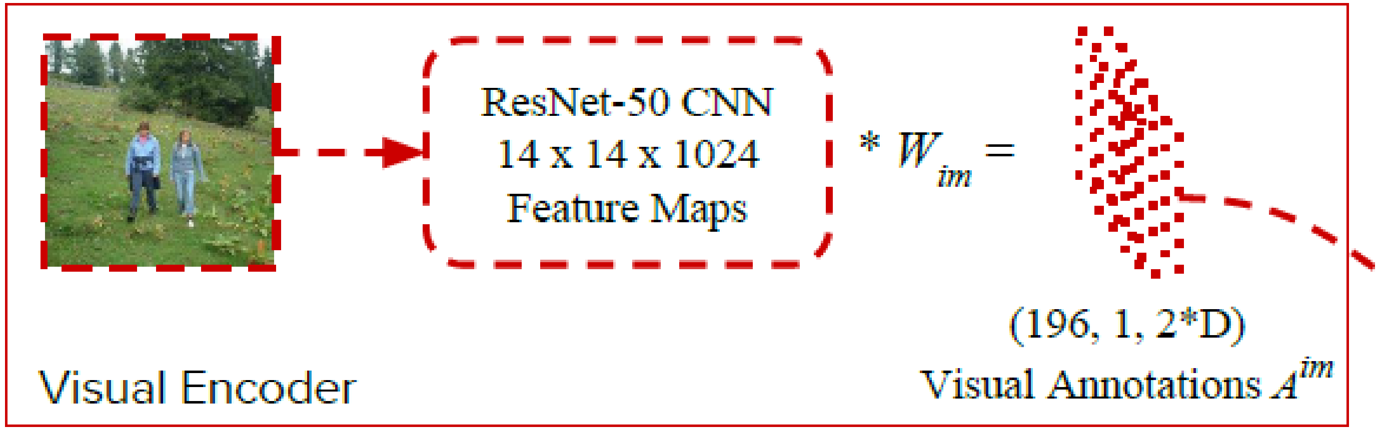
$$Y = (y_1, y_2, ..., y_M), y_j \in \mathbb{R}^E$$

bi-directional GRU

$$a_i^{txt} = \begin{bmatrix} \overrightarrow{h_i} \\ \overleftarrow{h_i} \end{bmatrix}, a_i^{txt} \in \mathbb{R}^{2D}$$

$$A^{txt} = \{a_1^{txt}, a_2^{txt}, \ldots, a_N^{txt}\}$$

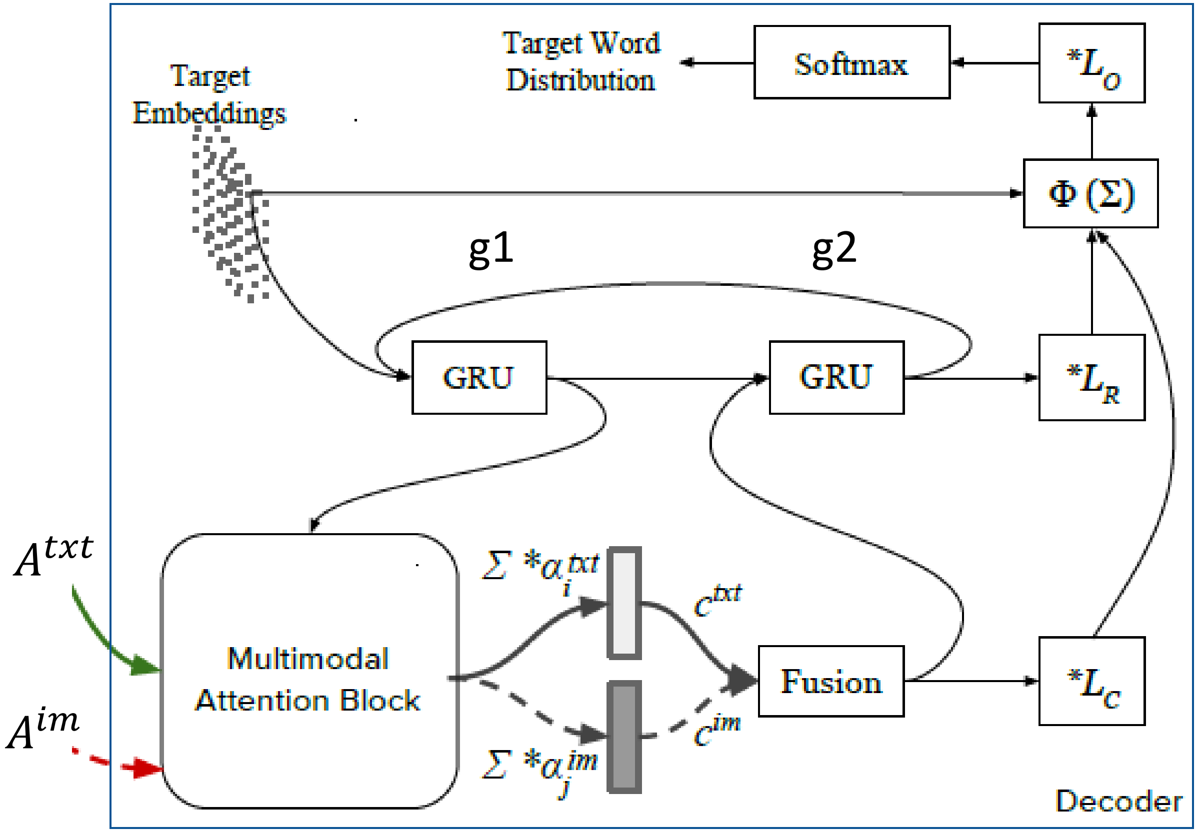
2.2.2 Visual Encoder



- Convolutional image features of size 14x14x1024 are extracted from the res4f_relu layer of ResNet-50 CNN trained on ImageNet.
- 196x1024 dimension per each image

$$A^{im} = \{a_1^{im}, a_2^{im}, \ldots, a_{196}^{im}\}$$

2.2.3 Decoder



$$A^{txt} = \{a_1^{txt}, a_2^{txt}, \ldots, a_N^{txt}\}$$

$$A^{im} = \{a_1^{im}, a_2^{im}, \ldots, a_{196}^{im}\}$$

$$h_0^{(1)} = tanh\left(W_{init}^T\left(\frac{1}{N}\sum_{i=1}^{N} a_i^{txt}\right) + b_{init}\right)$$

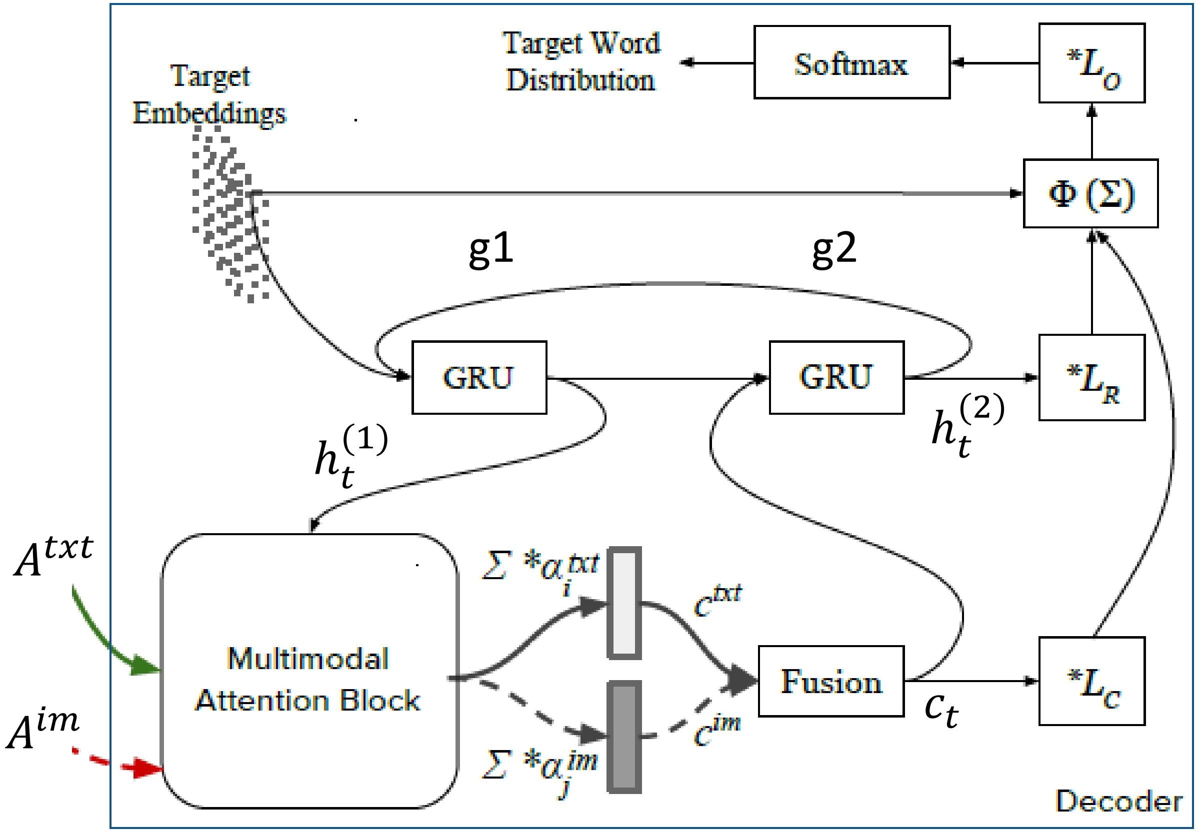
Init hidden state: learned from the mean textual annotation(A^{txt}) using a feed-forward layer with tanh nonlinearity.

$$h_0^{(2)} = h_1^{(1)}$$

At each time step, a multimodal attention mechanism computes two modality specific context vectors:

$$\{c_t^{txt}, c_t^{im}\}$$

2.2.3 Decoder (Cont'd)



$$\{A^{txt}, A^{im}\} \quad \{c_t^{txt}, c_t^{im}\}$$

Fusion: two different fusion techniques

- SUM
- CONCAT

$$c_t = F_S(c_t^{txt}, c_t^{im}) = \tanh(c_t^{txt} + c_t^{im})$$

$$c_t = F_C(c_t^{txt}, c_t^{im}) = \tanh\left(W_{fus}^T \begin{bmatrix} c_t^{txt} \\ c_t^{im} \end{bmatrix} + b_{fus}\right)$$

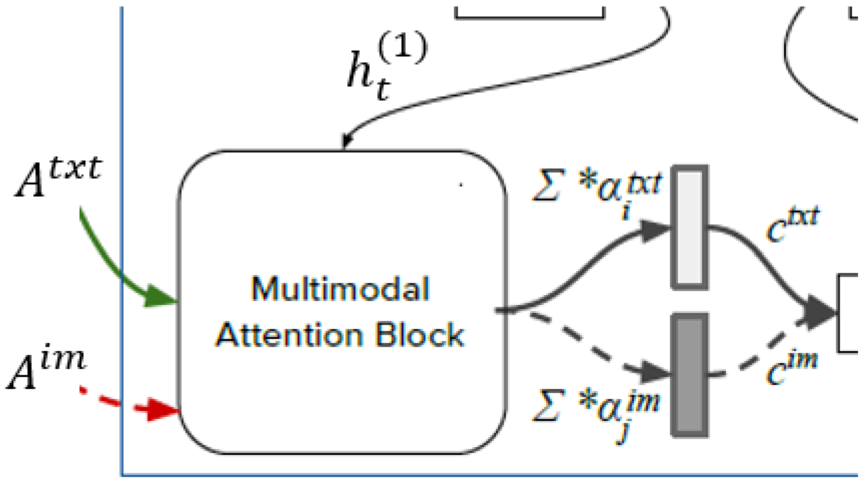
Probability distribution:

- hidden state $h_t^{(2)}$ of the second GRU
- multimodal context vector c_t
- the embedding of the (true) previous target word $E_{y_{t-1}}$

$$h_t^{(2)} = g_2(h_{t-1}^{(2)}, c_t)$$

$$P(y_t = k | y_{<t}, A^{txt}, A^{im}) = softmax\left(L_o \tanh(L_s h_t^{(2)} + L_c c_t + E_{y_{t-1}})\right)$$

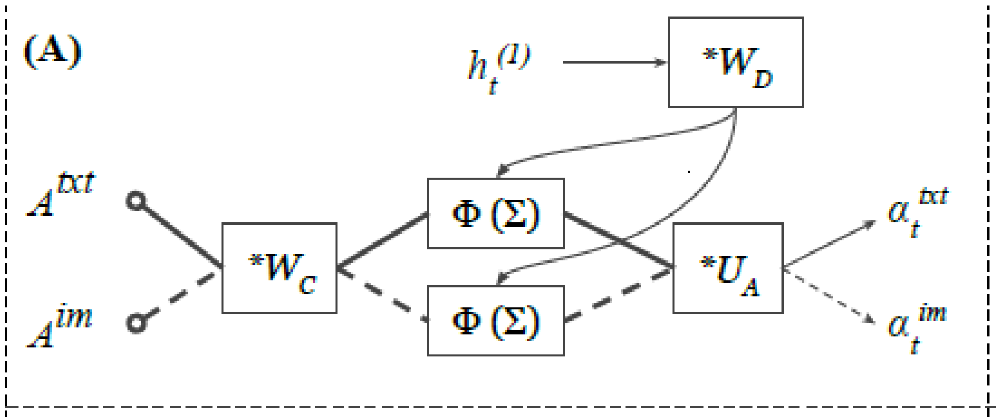
2.2.4 Attention Mechanism



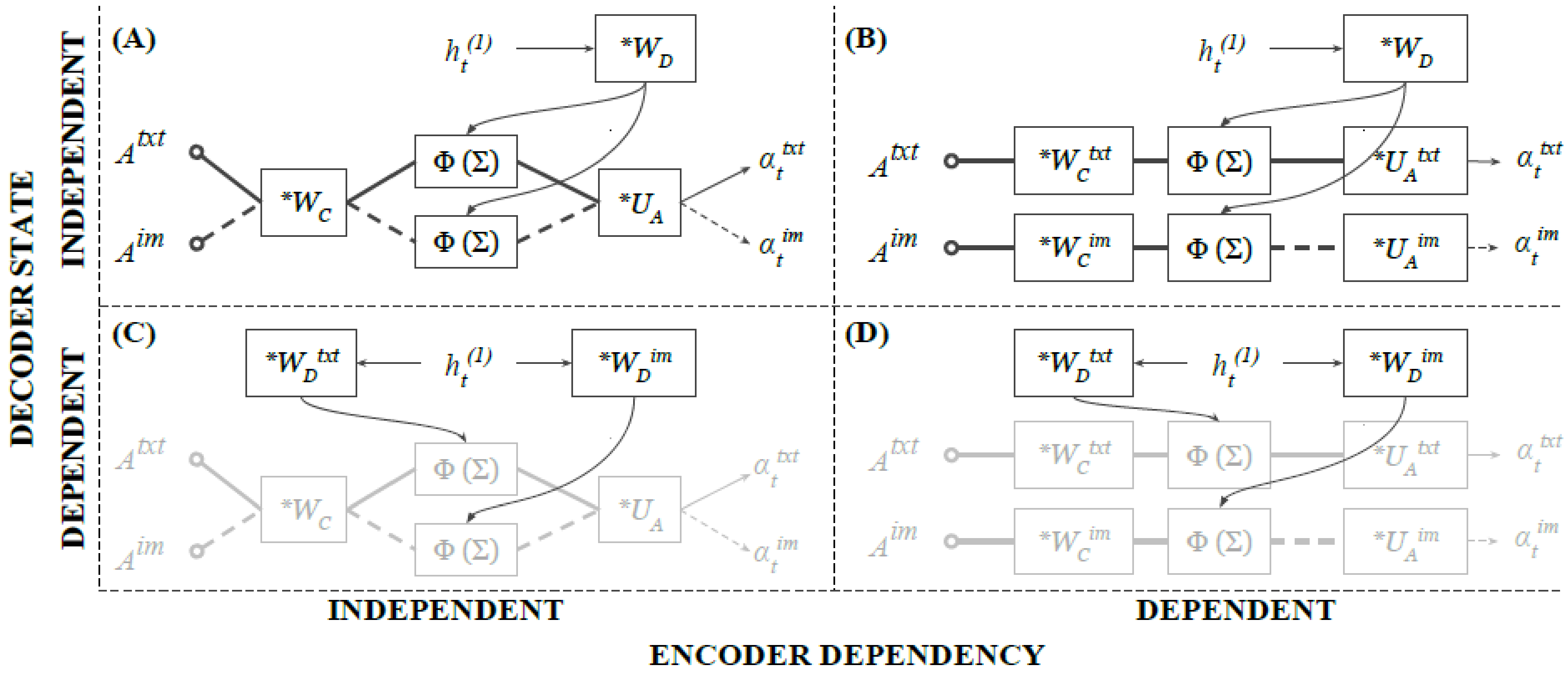
Use a shared feed-forward network:

$$\alpha_t^{txt} = softmax\left(U_A \tanh(W_D\, h_t^{(1)} + W_C\, A^{txt})\right)$$

$$\alpha_t^{im} = softmax\left(U_A \tanh(W_D\, h_t^{(1)} + W_C\, A^{im})\right)$$



2.2.4 Attention Mechanism (Cont'd)



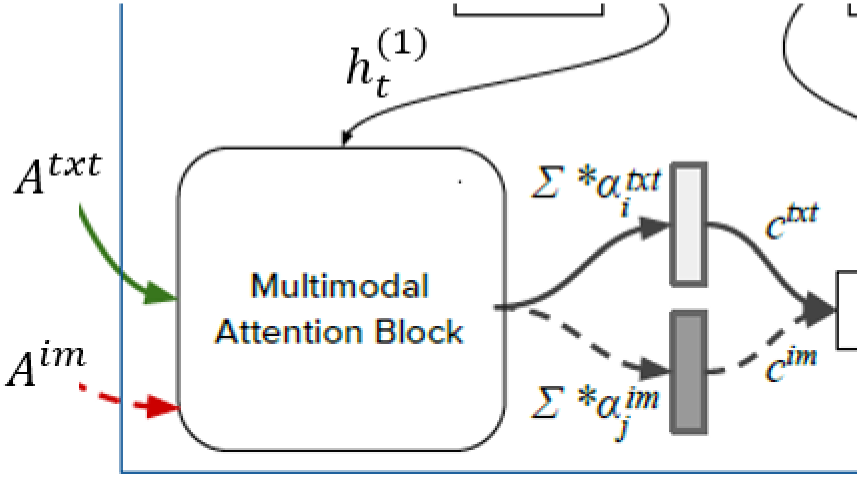
- (A)(C) encoder-independent
- (B)(D) encoder-dependent
- (A)(B) independent decoder state projection
- (C)(D) dependent decoder state projection

Four variants of the multimodal attention mechanism in terms of modality dependency with respect to encoder and decoder.

2.2.4 Attention Mechanism (Cont'd)

$$\{\alpha_t^{txt}, \alpha_t^{im}\}$$

$$c_t^{txt} = \sum_{i=1}^{N} \alpha_{ti}^{txt} \, a_i^{txt} \quad , \quad c_t^{im} = \sum_{j=1}^{196} \alpha_{tj}^{im} \, a_j^{im}$$



2.3 Experiment

	Attention Type			Validation Scores		
Model	Fusion	Modality	Decoder	METEOR	BLEU	CIDEr-D
NMT	-	-	-	34.24 (35.59)	18.64 (21.62)	58.57 (67.93)
IMGTXT	-	-	-	26.80	11.16	31.28
MNMT1	SUM	IND	IND	33.23 (35.42)	18.30 (21.24)	55.45 (65.03)
MNMT2	SUM	IND	DEP	34.17 (35.48)	17.70 (20.70)	53.78 (61.76)
MNMT3	SUM	DEP	IND	34.38 (35.55)	18.42 (20.94)	55.81 (63.37)
MNMT4	SUM	DEP	DEP	33.67 (34.57)	17.83 (20.30)	52.68 (59.63)
MNMT5	CONCAT	IND	IND	33.31 (34.98)	17.50 (20.60)	53.57 (61.46)
MNMT6	CONCAT	IND	DEP	**35.23** (36.79)	19.30 (22.45)	60.62 (69.96)
MNMT7	CONCAT	DEP	IND	**35.11 (37.13)**	**19.72 (23.24)**	**61.04 (72.16)**
MNMT8	CONCAT	DEP	DEP	34.80 (**36.98**)	19.55 (22.78)	60.20 (70.20)

Quantitative Analysis:

- SUM operator worse (MNMT1 – MNMT4)
  - concatenation makes use of a linear layer that learns how to integrate the modality-specific activations into the multimodal context vector.
- a completely independent (shared) attention mechanism (MNMT5) has the worst performance among all CONCAT variants.
  - different input modalities.
- Dependent encoder & independent decoder performs best.

2.3 Experiment (Cont'd)

Source		a woman in jeans and a red coat and carrying a multicolored handbag spreads put her arms while leaping up from a cobblestone street	an asian woman sitting on a bench going through a pink laptop bag	women in a black dress riding a scooter down the street
NMT	HYP	eine frau springt auf einem gehweg in die luft (47.84)	eine frau sitzt auf einer bank und hält einen laptop in der hand (52.61)	eine frau in weißem kleid fährt auf einem roller eine straße entlang (44.00)
	ENG	*a woman jumps on a walkway in the air*	*a woman sitting on a bench and* ***holding a laptop*** *in hand*	*a woman in a* ***white*** *dress riding a scooter a road along*
MNMT	HYP	eine frau in rotem anorak und schwarzer hose springt mit ausgebreiteten armen durch die luft (49.09)	eine asiatische frau sitzt auf einer holzbank (34.01)	eine frau im schwarzen kleid auf einem motorroller (31.15)
	ENG	*a woman in a* ***red suit and black trousers*** *jumping with* ***outstretched arms*** *through the air*	*an* ***asian*** *woman sitting on a* ***wooden bench***	*a woman in* ***black*** *dress on a scooter*

Qualitative Analysis

- Image 1, MNMT produces richer description
- Image 2, MNMT again produces rich description but ignores the pink laptop bag
- Image 3, NMT wrongly describes the color of an object while MNMT does its job correctly

2.4 Pros and cons

- Pros
  - Integrate natural language and image
  - Four variants of the multimodal attention mechanism

- Cons
  - How do image features enrich language information?
  - Why is the effect of the image unobvious?

Probing the Need for Visual Context in Multimodal Machine Translation

Ozan Caglayan
LIUM, Le Mans University
ozan.caglayan@univ-lemans.fr

Pranava Madhyastha
Imperial College London
pranava@imperial.ac.uk

Lucia Specia
Imperial College London
l.specia@imperial.ac.uk

Loïc Barrault
LIUM, Le Mans University
loic.barrault@univ-lemans.fr

1. Motivation

- Multimodal machine translation (MMT) has suggested that the visual modality is either unnecessary or only marginally beneficial.
- Current belief
  - MMT models disregard the visual modality because of either the quality of the image features or the way they are integrated into the model.
- Assumption
  - Natural language features is sufficient.
- Work
  - Probe the contribution of the visual modality to state-of-the-art MMT models by conducting a systematic analysis where we partially deprive the models from source-side textual context.

2. Method

2.1 Input Degradation

- Color Deprivation, $\mathcal{D}_C$
  - Replace source words that refer to colors with a special token [v]
- Entity Masking, $\mathcal{D}_N$
  - mask out the nouns in the source sentences with [v]
- Progressive Masking, $\mathcal{D}_k$
  - replaces all but the first k tokens of source sentences with [v].
- Visual Sensitivity
  - Feed the visual features in reverse sample order to break image-sentence alignments in test-time.

2.1 Input Degradation (Cont'd)

$\mathcal{D}$	a	lady	in	a	blue	dress	singing

- Color Deprivation, $\mathcal{D}_C$

$\mathcal{D}_C$	a	lady	in	a	[v]	dress	singing

- Entity Masking, $\mathcal{D}_N$

$\mathcal{D}_N$	a	[v]	in	a	blue	[v]	singing

- Progressive Masking, $\mathcal{D}_k$

$\mathcal{D}_4$	a	lady	in	a	[v]	[v]	[v]
$\mathcal{D}_2$	a	lady	[v]	[v]	[v]	[v]	[v]
$\mathcal{D}_0$	[v]	[v]	[v]	[v]	[v]	[v]	[v]

- Visual Sensitivity

2.2 Models

- NMT (neural machine translation)

- DIRECT (basic multimodal attention)
  - Linearly projects the concatenation of textual and visual context vectors to obtain the multimodal context vector

- HIER (hierarchical extension of DIRECT)
  - while the latter replaces the concatenation with another attention layer

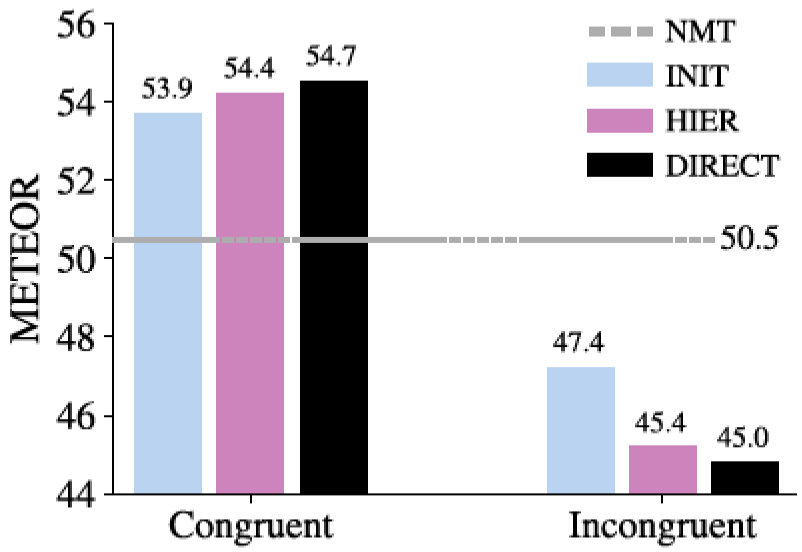
- INIT
  - initialize both the encoder and the decoder using a non-linear transformation of the pool5 features.

2.3 Experiment

Color Deprivation, $\mathcal{D}_C$

	$\mathcal{D}$	$\mathcal{D}_C$
NMT	70.6 ± 0.5	68.4 ± 0.1
INIT	70.7 ± 0.2	$\mathbf{68.9} \pm 0.1$
HIER	70.9 ± 0.3	$\mathbf{69.0} \pm 0.3$
DIRECT	70.9 ± 0.2	$\mathbf{68.8} \pm 0.3$

Entity Masking, $\mathcal{D}_N$



- The gains are much more prominent
- Visual modality is now much more important with entity masking.

2.3 Experiment

Entity Masking, $\mathcal{D}_N$

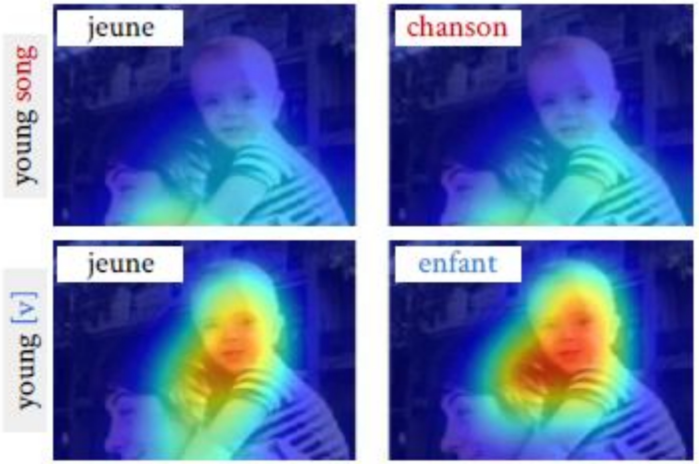


Figure 2: Baseline MMT (top) translates the misspelled "son" while the masked MMT (bottom) correctly produces "enfant" (child) by focusing on the image.

the masked MMT model attends
to the correct region of the image

Entity masking results across three languages

	+ Gain (↓ Incongruence Drop)		
	INIT	HIER	DIRECT
Czech	+1.4 (↓ 2.9)	+1.7 (↓ 3.5)	+1.7 (↓ 4.1)
German	+2.1 (↓ 4.7)	+2.5 (↓ 5.9)	+2.7 (↓ 6.5)
French	+3.4 (↓ 6.5)	+3.9 (↓ 9.0)	+4.2 (↓ 9.7)

Table 3: *Entity masking* results across three languages: all MMT models perform significantly better than their NMT counterparts (p-value ≤ 0.01). The incongruence drop applies on top of the MMT score.

2.3 Experiment

Progressive Masking, $\mathcal{D}_k$

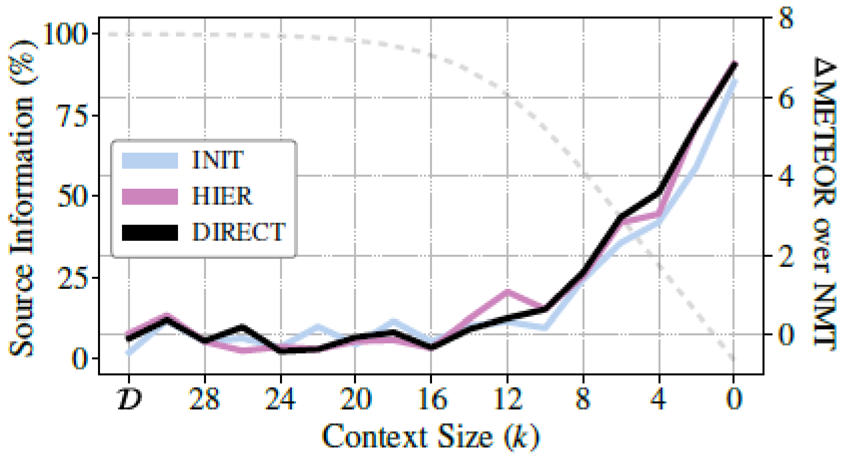


Figure 3: Multimodal gain in absolute METEOR for *progressive masking*: the dashed gray curve indicates the percentage of non-masked words in the training set.

2.3 Experiment

Qualitative examples

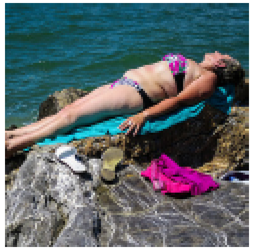
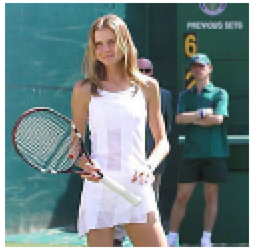
	SRC:	an older woman in [v] [v] [v] [v] [v] [v] [v] [v] [v] [v] [v]
	NMT:	une femme âgée avec un t-shirt blanc et des lunettes de soleil est assise sur un banc
		(an older woman with a white t-shirt and sunglasses is sitting on a bank)
	MMT:	une femme âgée en **maillot de bain rose** est assise sur un **rocher au bord de l'eau**
		(an older woman with a pink swimsuit is sitting on a rock at the seaside)
	REF:	une femme âgée **en bikini** bronze sur **un rocher au bord de l'océan**
		(an older woman in bikini is tanning on a rock at the edge of the ocean)
	SRC:	a young [v] in [v] holding a tennis [v]
	NMT:	un jeune garçon en bleu tenant une raquette de tennis
		(a young boy in blue holding a tennis racket)
	MMT:	**une** jeune **femme** en **blanc** tenant une raquette de tennis
	REF:	**une** jeune **femme** en **blanc** tenant une raquette de tennis
		(a young girl in white holding a tennis racket)
	SRC:	little girl covering her face with a [v] towel
	NMT:	une petite fille couvrant son visage avec une serviette blanche
		(a little girl covering her face with a white towel)
	MMT:	une petite fille couvrant son visage avec une serviette **bleue**
	REF:	une petite fille couvrant son visage avec une serviette **bleue**
		(a little girl covering her face with a blue towel)

Table 5: Qualitative examples from progressive masking, entity masking and color deprivation, respectively. Underlined and bold words highlight the bad and good lexical choices. MMT is an attentive system.

2.4 Pros and cons

- Pros
  - Sufficient experiment
  - In-depth study on the contribution of images for multimodal machine translation.
  - Models are able to integrate the visual modality if the available modalities are complementary rather than redundant.

- Cons
  - How to use this conclusion to get better multimodal machine translation result.

2.5 Inspiration

- How to add multimodality to our own work
- In-depth analysis of predecessors' work