Minneapolis, MN
NAACL
HLT 2019
June 2-7, 2019

Probing the Need for Visual Context in Multimodal Machine Translation

Ozan Caglayan[1], Pranava Madhyastha[2], Lucia Specia[2], Loïc Barrault[1]

[1] Le Mans Université
[2] Imperial College London

Multimodal Machine Translation (MMT)

- Better machine translation approaches by leveraging multiple **modalities**
- Dataset → Multi30K (Elliott et al., 2016)
  - Multilingual extension of Flickr30K (Young et al., 2014)
  - **Images**, **English** descriptions, **French**, **German** and **Czech** translations.

Potential benefit

- Language grounding
  - Sense disambiguation → “river bank” vs. “financial bank”
  - Grammatical gender disambiguation
  - Learning concepts

Example: grammatical gender

Source Sentence (EN)

A baseball player in a black shirt just tagged **a player** in a white shirt.

Candidate Translations (FR)

Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.

"Female" baseball player

Un joueur de baseball en maillot noir vient de toucher **un joueur** en maillot blanc.

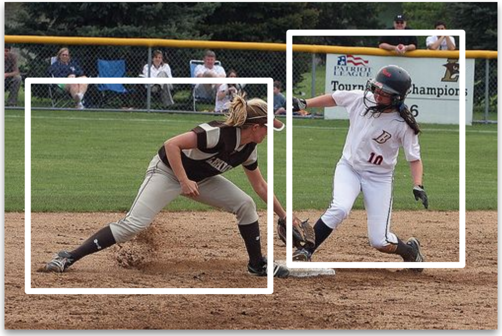
"Male" baseball player

Example: grammatical gender

Visual context disambiguates the gender

Source Sentence (EN)

A baseball player in a black shirt just tagged **a player** in a white shirt.



Candidate Translations (FR)

✓ **Une joueuse** de baseball en maillot noir vient de toucher **une joueuse** en maillot blanc.

"Female" baseball player

Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.

"Male" baseball player

Where are we?

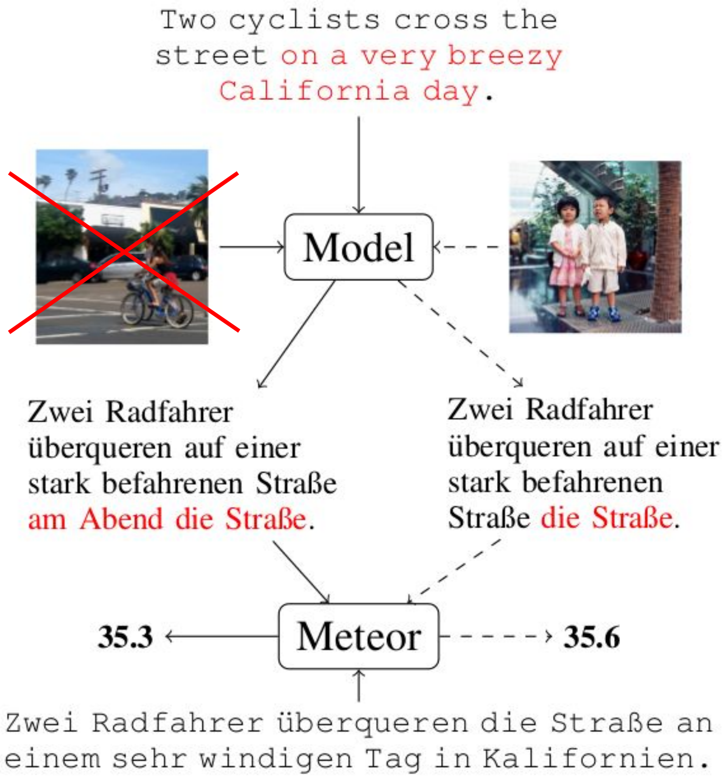
- Benefit of current approaches is not evident - WMT18 (Barrault et al., 2018):
  - Largest gain from external corpora, not from images (Grönroos et al., 2018)

EN → DE	BLEU ↑	Meteor ↑
•MeMAD_1_FLICKR_DE_MeMAD-OpenNMT-mmod_U (P)	**38.5**	**56.6**
CUNI_1_FLICKR_DE_NeuralMonkeyTextual_U	32.5	52.3
CUNI_1_FLICKR_DE_NeuralMonkeyImagination_U (P)	32.2	51.7
UMONS_1_FLICKR_DE_DeepGru_C (P)	31.1	51.6
LIUMCVC_1_FLICKR_DE_NMTEnsemble_C (P)	31.1	51.5
LIUMCVC_1_FLICKR_DE_MNMTEnsemble_C (P)	31.4	51.4
OSU-BD_1_FLICKR_DE_RLNMT_C (P)	32.3	50.9
SHEF_1_DE_MLT_C (P)	30.4	50.7
SHEF1_1_DE_MFS_C (P)	30.3	50.7
Baseline	27.6	47.4
AFRL-OHIO-STATE_1_FLICKR_DE_4COMBO_U (P)	24.3	45.4

Where are we?

- Benefit of current approaches is not evident:
    - Adversarially attacking MMT marginally influences the scores (Elliott 2018)

METEOR (EN-DE)	Congruent	Incongruent
Dec-init	57.0	56.8
Trg-mul	57.3	57.3
Fusion-conv	55.0	53.3



(b) Incongruent is better than Congruent

Why don't images help?

- Pre-trained CNN **features** may not be good enough for MMT
    - ImageNet has very limited set of objects

- Current **multimodal models** may not be effective

- Multi30K **dataset** may be
    - Too simple; language is enough
    - Too small to generalise visual features

Why don't images help?

- Pre-trained CNN **features** may not be good enough for MMT
  - ImageNet has very limited set of objects
- Current **multimodal models** may not be effective
- Multi30K **dataset** may be
  - Too simple; language is enough
  - Too small to generalise visual features

This paper

- We **degrade source language**
  - Systematically mask source words at **training** and **inference** times

- **Hypothesis 1**: **MMT models** should perform better than **text-only models** if image is effectively taken into account
  - Image features ✓
  - Multimodal models ✓

- **Hypothesis 2**: More **sophisticated MMT models** should perform better than **simpler MMT models**

Types of degradation

Source sentence **"a lady in a blue dress singing"**

Types of degradation (1)

Source sentence **"a lady in a blue dress singing"**

Color Masking	a	lady	in	a	[v]	dress	singing

- Very small-scale masking
    - **3.3%** of source words are removed

Types of degradation (2)

Source sentence **“a lady in a blue dress singing”**

Color Masking	a	lady	in	a	[v]	dress	singing
Entity Masking	a	[v]	in	a	blue	[v]	singing

- Uses Flickr30K entity annotations (Plummer et al., 2015)
  - **26%** of source words are removed (3.4 blanks / sent)

Types of degradation (3)

Source sentence **"a lady in a blue dress singing"**

Color Masking	a	lady	in	a	[v]	dress	singing
Entity Masking	a	[v]	in	a	blue	[v]	singing
Progressive Masking (k=4)	a	lady	in	a	[v]	[v]	[v]
Progressive Masking (k=2)	a	lady	[v]	[v]	[v]	[v]	[v]
Progressive Masking (k=0)	[v]	[v]	[v]	[v]	[v]	[v]	[v]

- Removal of **any** words
  - 16 variants with $k \in \{0, 2, \ldots, 30\}$
  - MMT task becomes multimodal sentence completion/captioning

Settings

- 2-layer GRU-based encoder/decoder NMT
  - 400D hidden units, 200D embeddings

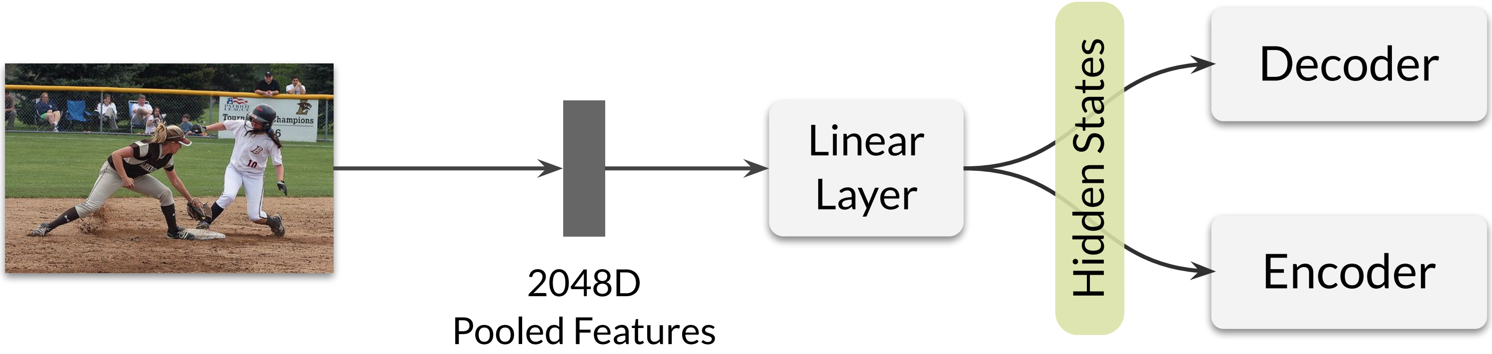
- Visual features → ResNet-50 CNN pretrained on ImageNet
  - 2048D pooled vectoral representations
  - 2048x8x8 convolutional feature maps

- Multi30K dataset
  - Primary language pair: **English** → **French**

MMT methods

Simple grounding

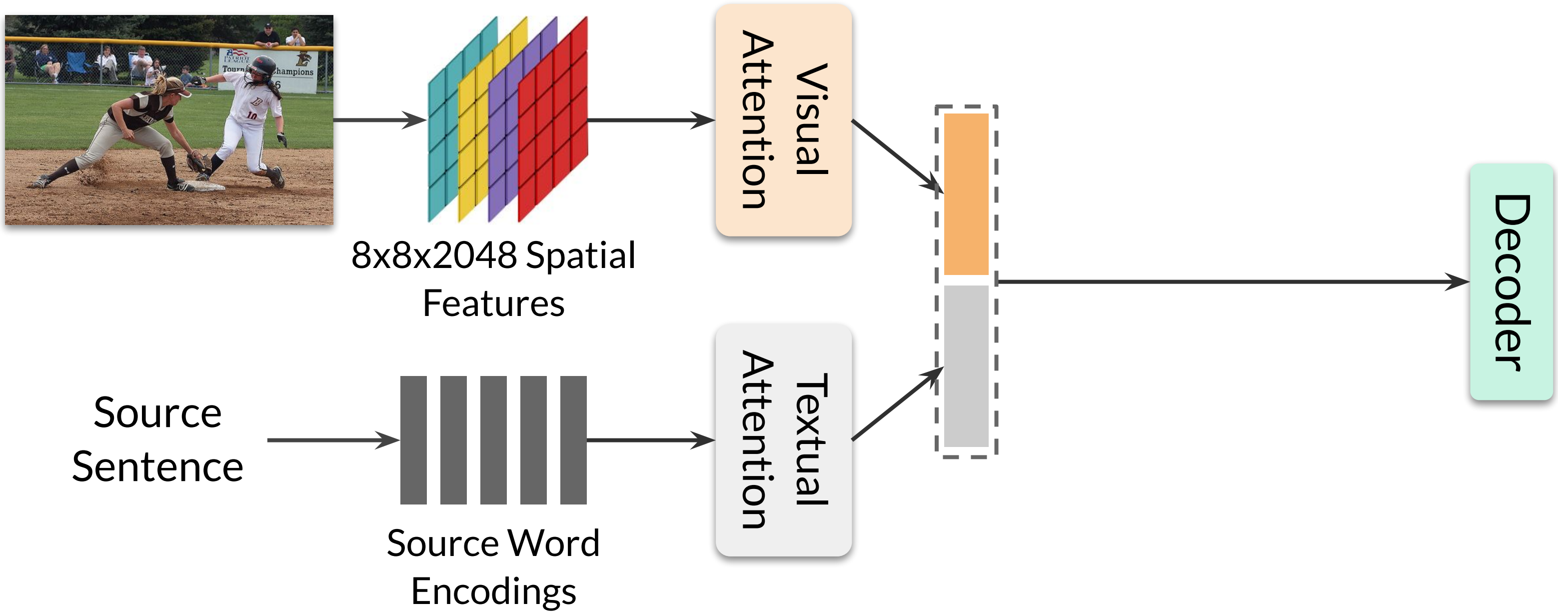
- **Tied INITialization** of encoders and decoders
  (Calixto and Liu, 2017), (Caglayan et al., 2017)



MMT methods

Multimodal attention

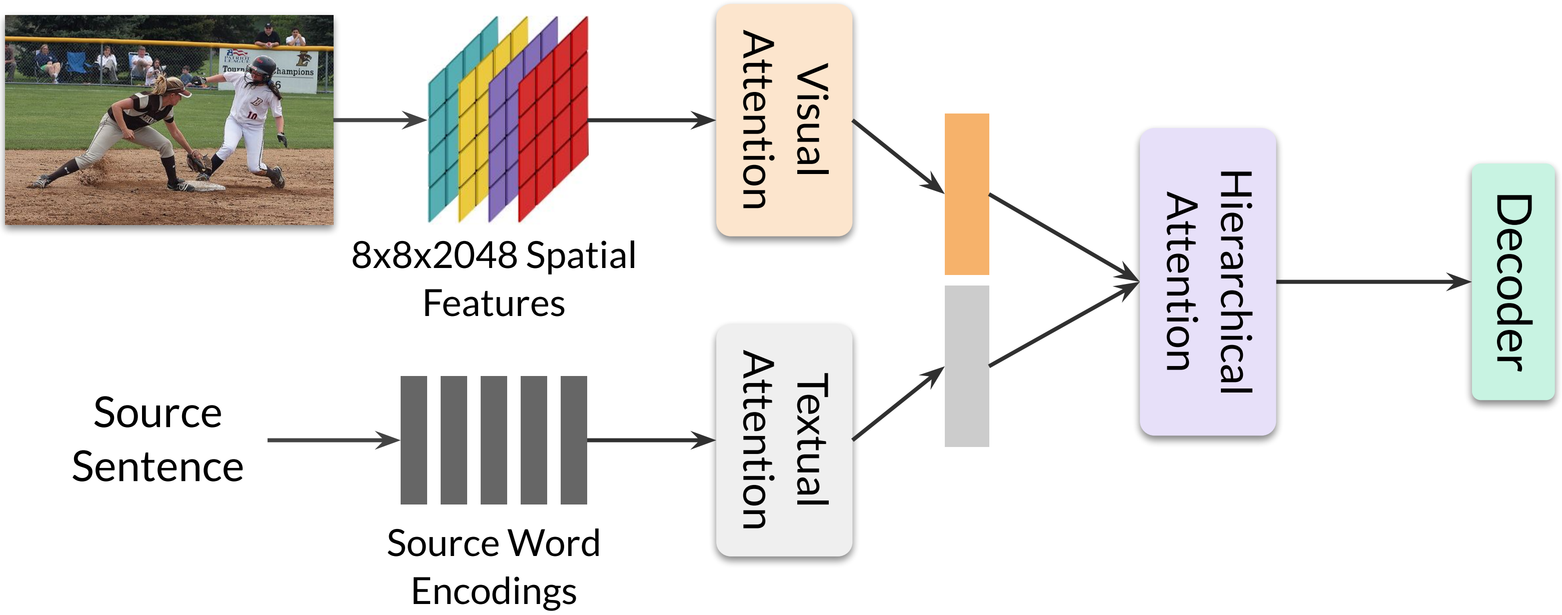
- **DIRECT fusion** uses modality specific attention layers and concatenates their output (Caglayan et al., 2016), (Calixto et al., 2016)



MMT methods

Multimodal attention

- **HIERarchical fusion** applies a third attention layer instead of concatenation (Libovický and Helcl, 2017)



Evaluation

- Mean and standard deviation (3 runs) of METEOR scores
- Statistical significance testing with `MultEval` (Clark et al., 2011)

Adversarial evaluation → Shuffled (**incongruent**) image features (Elliott 2018)

- **Incongruent decoding:** Incongruent features at inference time-only
- **Blinding:** Incongruent features at training and inference times

Results

Upper bound - no masking

Method	Baseline METEOR
NMT	70.6 ± 0.5
INIT	70.7 ± 0.2
HIER	70.9 ± 0.3
DIRECT	70.9 ± 0.2

- MMTs slightly better than NMT on average

Color masking

Method	Baseline METEOR	Masked METEOR
NMT	70.6 ± 0.5 →	68.4 ± 0.1
INIT	70.7 ± 0.2	
HIER	70.9 ± 0.3	
DIRECT	70.9 ± 0.2	

- Masked NMT suffers a substantial 2.2 drop

Color masking

Method	Baseline METEOR	Masked METEOR
NMT	70.6 ± 0.5	68.4 ± 0.1
INIT	70.7 ± 0.2	68.9 ± 0.1
HIER	70.9 ± 0.3	69.0 ± 0.3
DIRECT	70.9 ± 0.2	68.8 ± 0.3

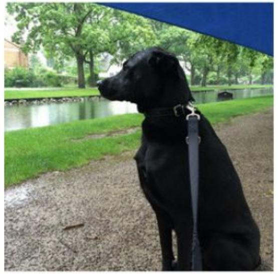
- Masked NMT suffers a substantial 2.2 drop
- Masked MMT significantly better than masked NMT

Color masking

Method	Baseline METEOR	Masked METEOR	Masked color Accuracy (%)
NMT	70.6 ± 0.5	→ 68.4 ± 0.1	32.5
INIT	70.7 ± 0.2	68.9 ± 0.1	36.5
HIER	70.9 ± 0.3	69.0 ± 0.3	44.5
DIRECT	70.9 ± 0.2	68.8 ± 0.3	44.5

- Masked NMT suffers a substantial 2.2 drop
- Masked MMT significantly better than masked NMT
- **Accuracy** in color translation much better in attentive MMT

Color masking

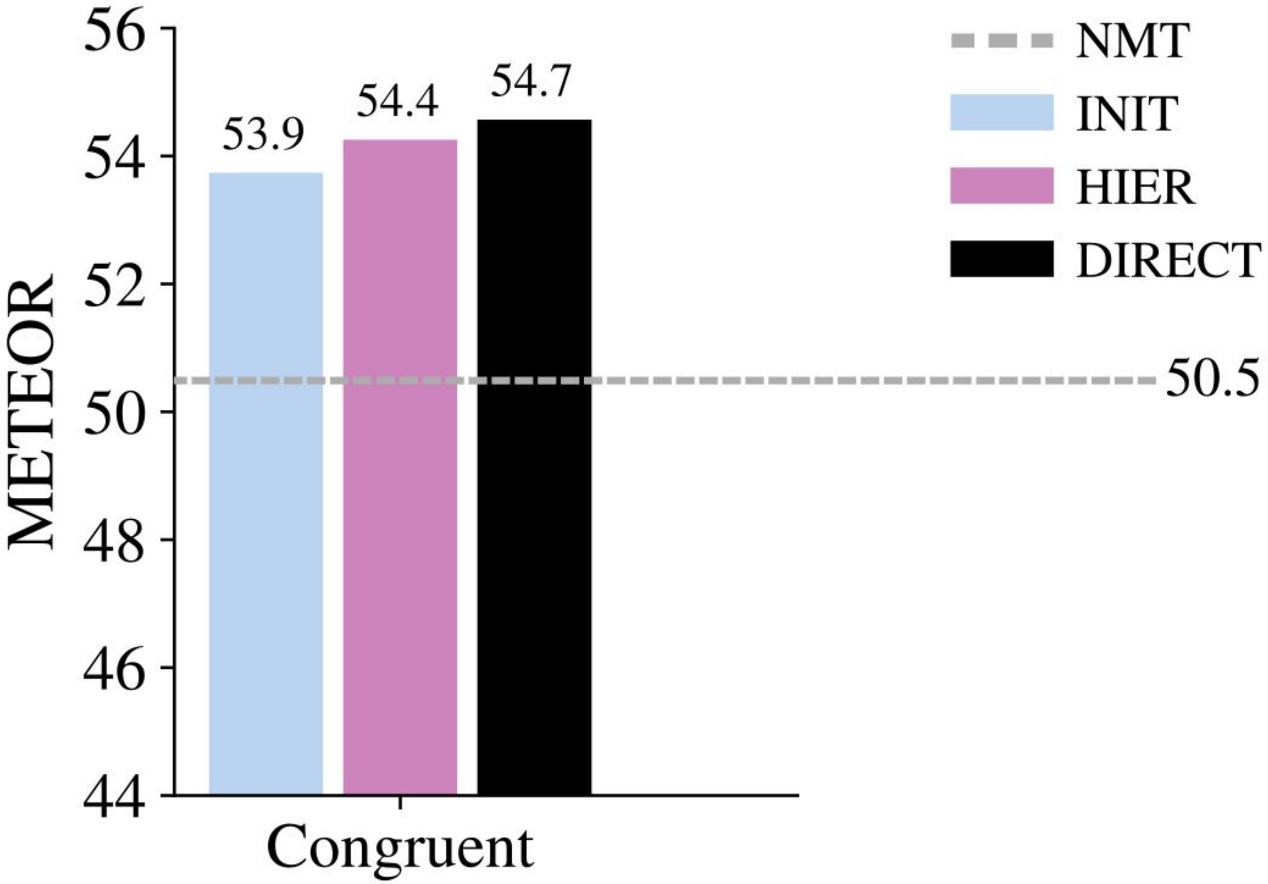


SRC: a [v] dog sits under a [v] umbrella
NMT: brown / **blue**
Init: **black / blue**
Hier: **black / blue**
Direct: **black / blue**

SRC: a woman in a [v] top is dancing as a woman and boy in a [v] shirt watch
NMT: blue / blue
Init: blue / blue
Hier: **red / red**
Direct: **red / red**

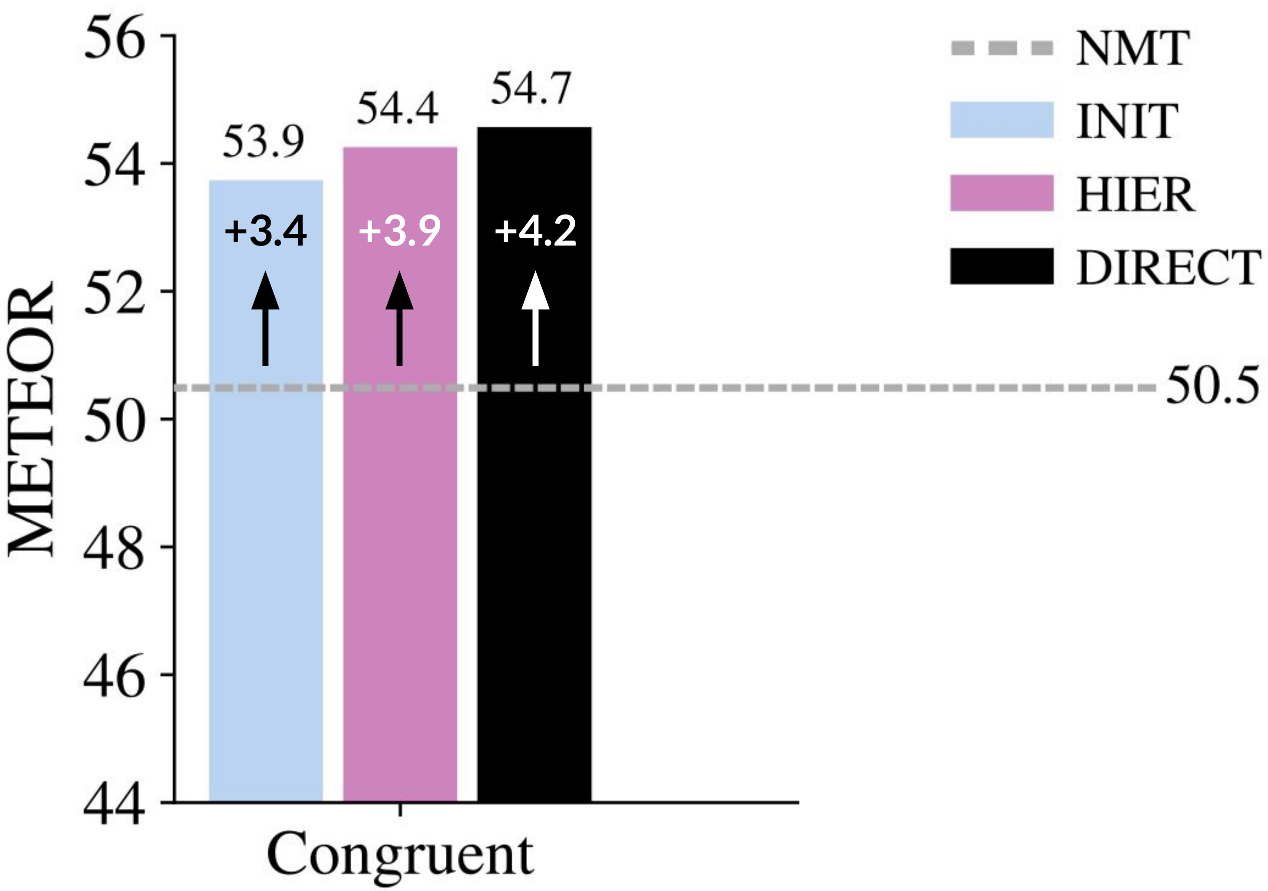
SRC: three female dancers in [v] dresses are performing a dance routine
NMT: white
Init: white
Hier: white
Direct: **blue**

Entity masking



NMT suffers > 20 points drop

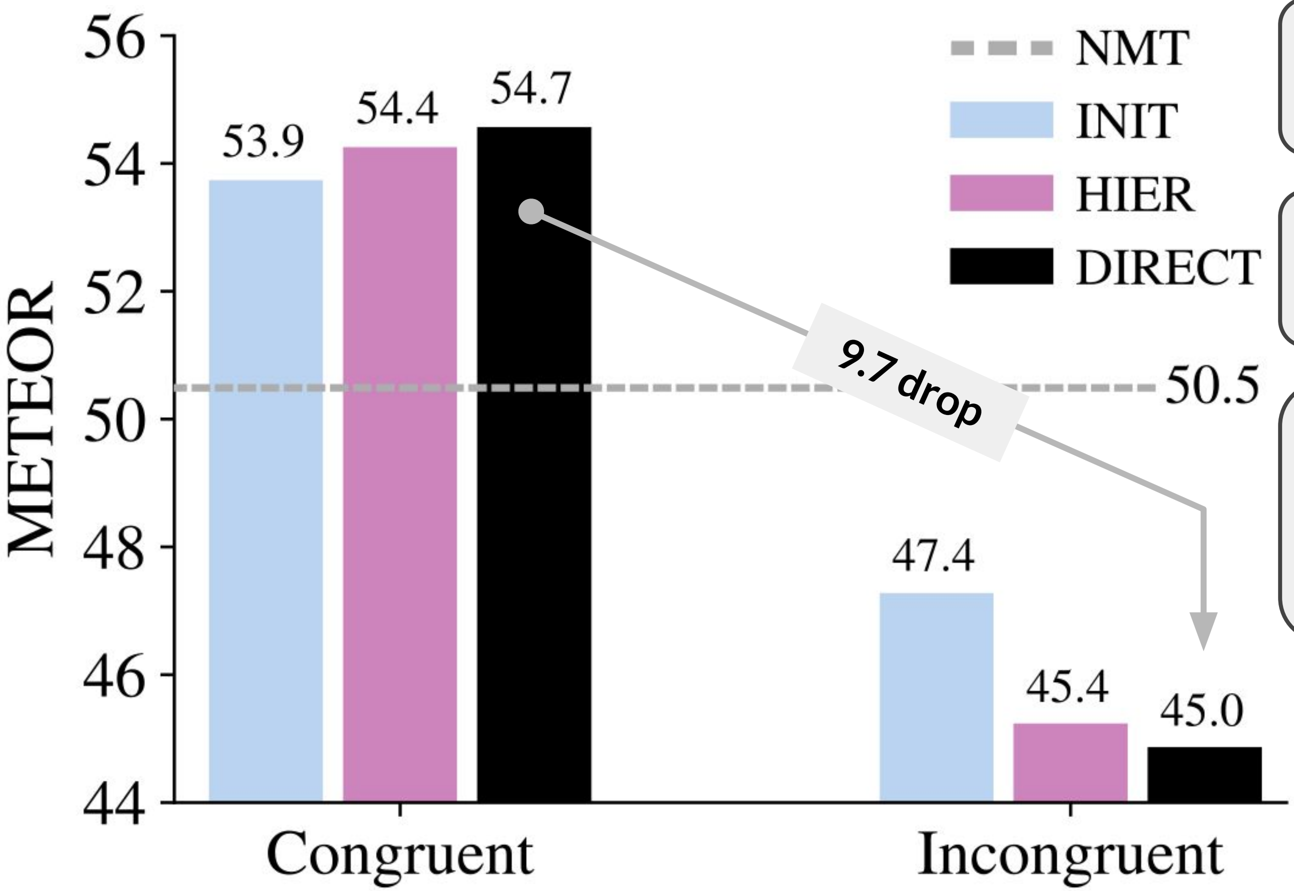
Entity masking



NMT suffers > 20 points drop

Up to 4.2 METEOR recovered by MMT

Entity masking



NMT suffers > 20 points drop

Up to 4.2 METEOR recovered by MMT

Models are visually sensitive: Up to ~**10 METEOR** drop with incongruent decoding

Entity masking (all languages)

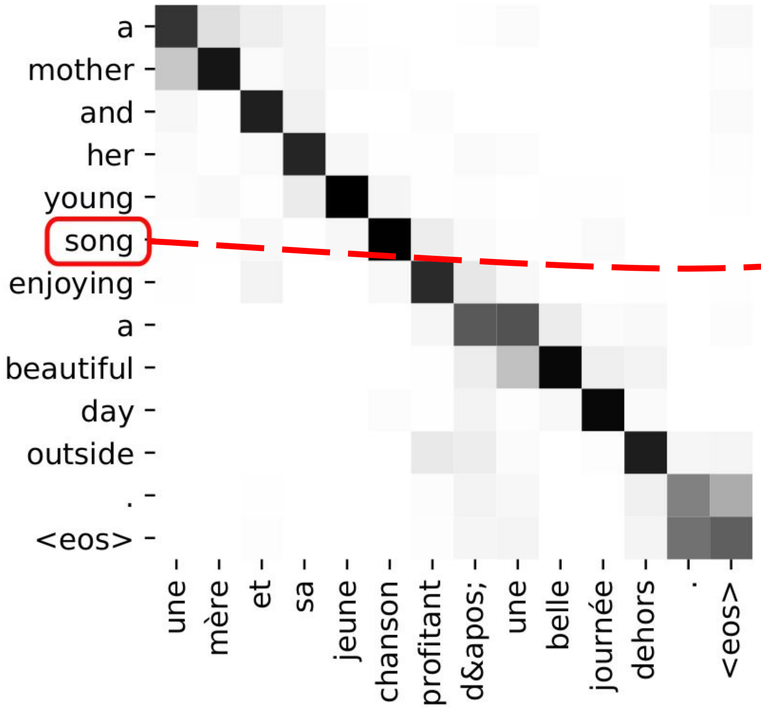
English →	MMT Gain over NMT			
	INIT	HIER	DIRECT	Average
Czech	1.4	1.7	1.7	1.6
German	2.1	2.5	2.7	2.4
French	3.4	3.9	4.2	**3.8**
Average	2.3	2.7	**2.9**	

All languages benefit from visual context

French benefits the most (less morphology)

Multimodal **attention** better than **INIT, Direct** fusion slightly better than **hierarchical**

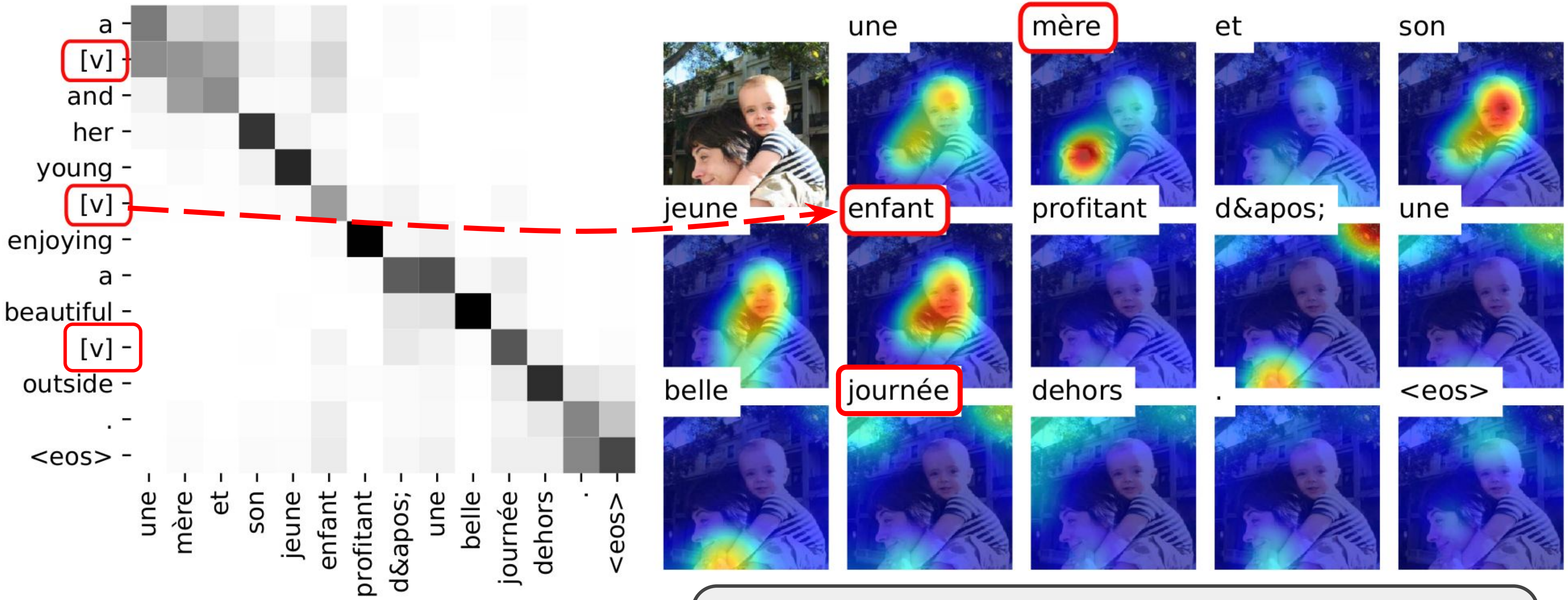
Entity masking (attention)



A typo in the source (**song**) - translated to "**chanson**"

Visual attention barely changes

Entity masking (attention)



“**mother**”, “**song**” and “**day**” are masked

Textual attention is less confident, visual attention works!

Entity masking

MMT is attentive, **INC** is incongruent decoding

SRC: a [v] drinks [v] outside on the [v]
REF: a dog drinks water outside on the grass

NMT: a man drinks wine outside on the sidewalk
MMT: a dog drinks water outside on the grass
INC: a man drinks flowers outside on the grass

SRC: a [v] turns on the [v] to pursue a flying [v]
REF: a dog turns on the grass to chase a ball in the air

NMT: a man turns on the beach to catch a flying frisbee
MMT: a dog turns on the grass to catch a flying frisbee
INC: a woman turns around on the sidewalk to make a flying object

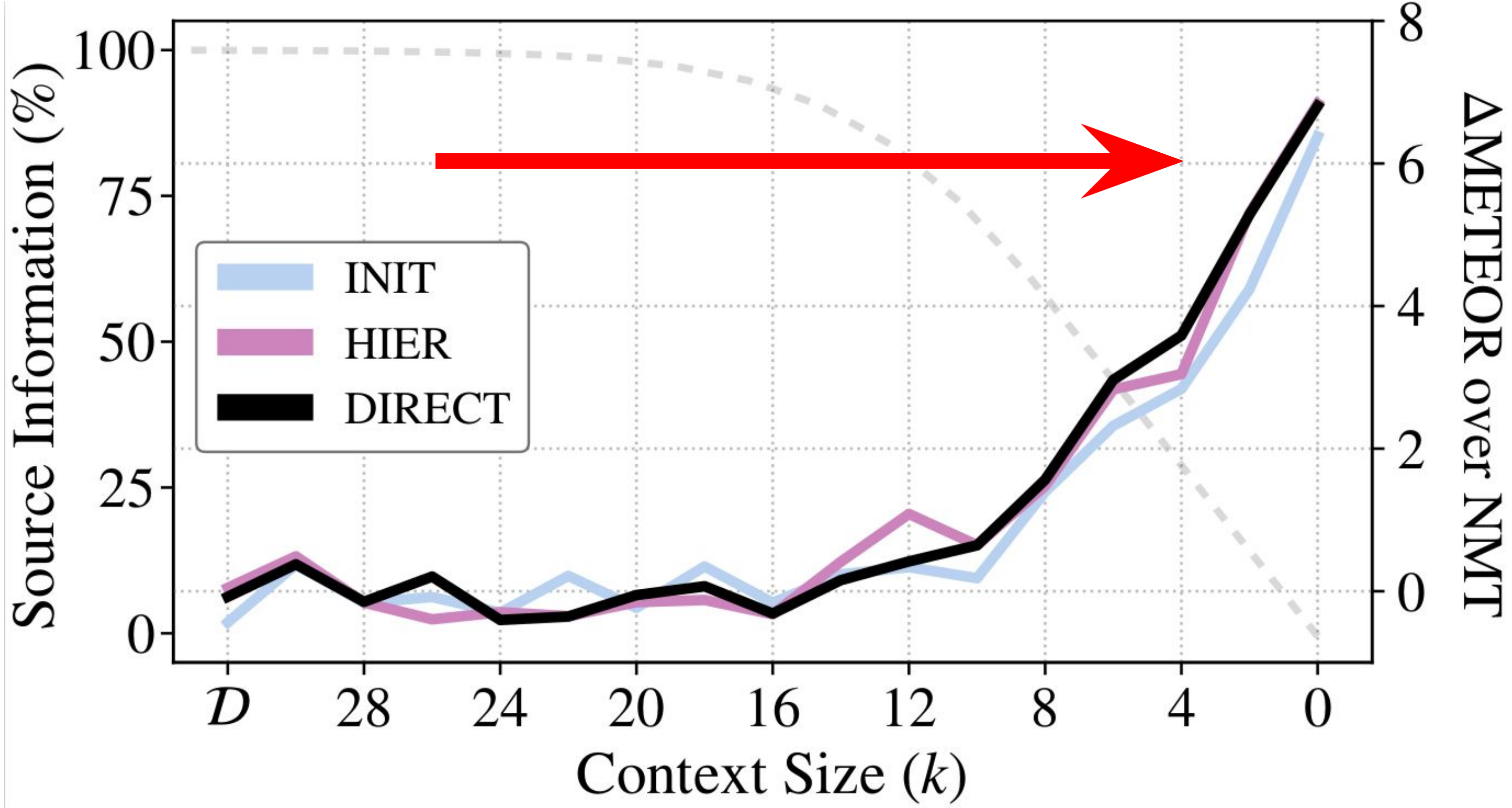
---

SRC: a young [v] in [v] holding a tennis [v]
REF: a young girl in white holding a tennis racket

NMT: a young boy in blue holding a tennis racket
MMT: a young girl in white holding a tennis racket
INC: a young man in blue holding a tennis ball

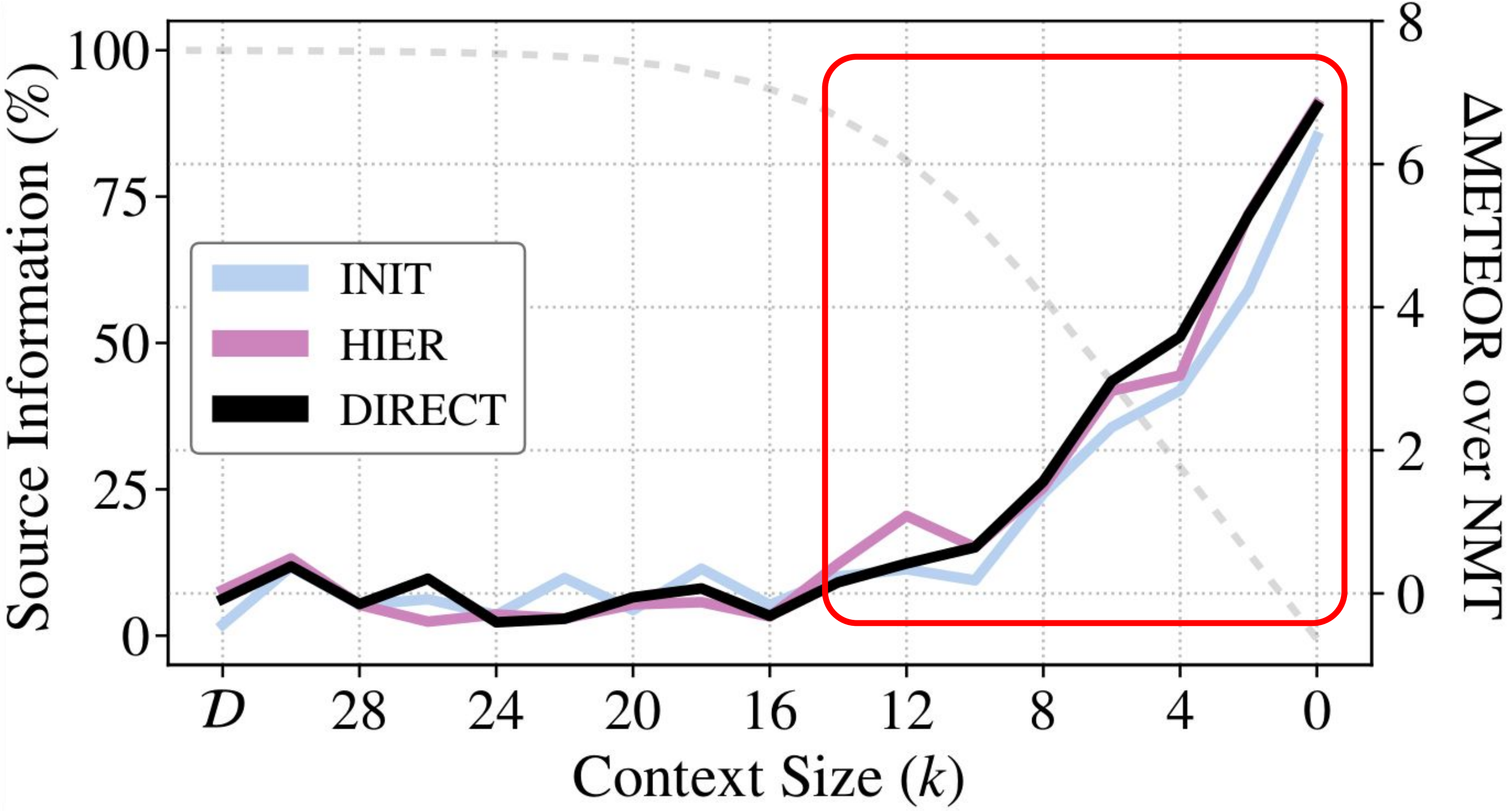
Progressive masking

As more information is removed, all MMT models leverage visual context, up to 7 METEOR points



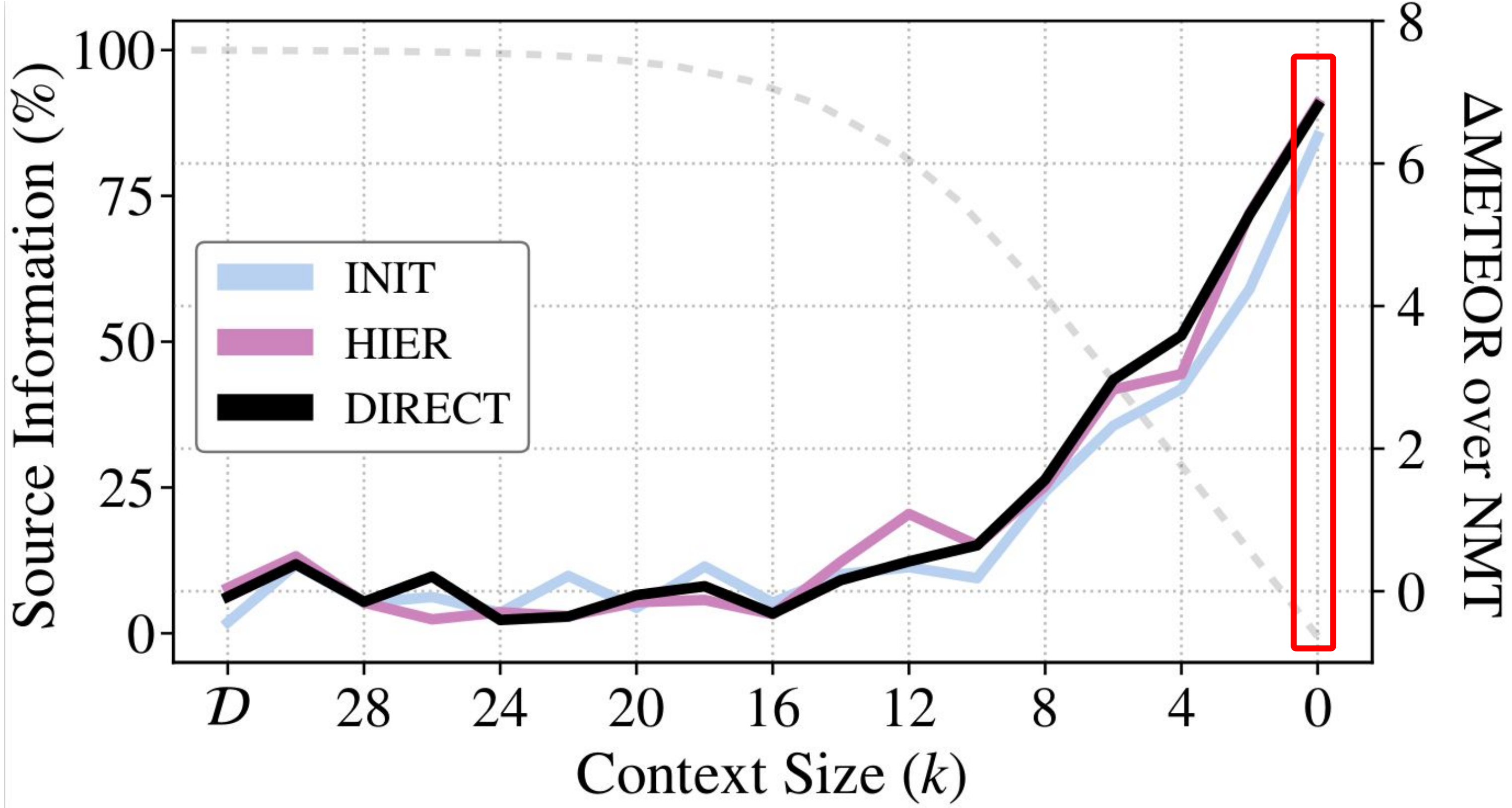
Progressive masking

Attentive models perform better than **INIT**



Progressive masking

Upper bound: ~7 METEOR when all words are masked



Progressive masking

	Original	k=12	k=4
NMT	70.6	63.9	28.6

- Compare two degraded variants to original Multi30K

Progressive masking

	Original	k=12	k=4
NMT	70.6	63.9	28.6
DIRECT MMT	+ 0.3	+ 0.6	+ 3.7

- Compare two degraded variants to original Multi30K
- MMT improves over NMT as linguistic information (k) is removed

Progressive masking

	Original	k=12	k=4	
NMT	70.6	63.9	28.6	
DIRECT MMT	+ 0.3	+ 0.6	+ 3.7	
Incongruent Dec.	- 0.7	- 1.4	- 6.4	(Relative to DIRECT MMT)

- Compare two degraded variants to original Multi30K
- MMT improves over NMT as linguistic information (k) is removed
  - It also becomes sensitive to the visual incongruence

Progressive masking

	Original	k=12	k=4	
NMT	70.6	63.9	28.6	
DIRECT MMT	+ 0.3	+ 0.6	+ 3.7	
Incongruent Dec.	- 0.7	- 1.4	- 6.4	(Relative to DIRECT MMT)
Blinding	70.6	64.1	28.4	

- Compare two degraded variants to original Multi30K
- MMT improves over NMT as linguistic information (k) is removed
  - It also becomes sensitive to the visual incongruence
- MMT that **never** sees correct features converges to text-only NMT
  - MMT improvements are not random

Progressive masking

MMT is attentive, **INC** is incongruent decoding

SRC: trees are in front [v][v][v][v][v]
REF: trees are in front of a big mountain

NMT: bicycles are in front of an outdoor building
MMT: **trees are in front of the mountain**
INC: taxis are in front of the window of a car

SRC: girls wave purple flags [v][v][v][v][v][v][v]
REF: girls wave purple flags as they parade down the street

NMT: girls in purple t-shirts are sitting on chairs in a classroom
MMT: **girls in purple costumes dance on a city street**
INC: girls in red shirts riding a bicycle in a city street

SRC: an older woman in [v][v][v][v][v][v][v][v][v][v][v]
REF: an older woman in bikini is tanning on a rock at the edge of the ocean

NMT: an older woman with a white t-shirt and sunglasses is sitting on a bank
MMT: **an older woman with a pink swimsuit is sitting on a rock at the seaside**
INC: an older woman in white t-shirt is standing next to a large tree

Conclusion

- **Hypothesis 1**: **MMT models** should perform better than **text-only models** if image is effectively taken into account
  - Visual info is taken into account if modalities are **complementary** rather than **redundant**
  - **Incorrect** visual info harms performance substantially more

- **Hypothesis 2**: More **sophisticated MMT models** should perform better than **simpler MMT models**
  - Attentive MMT better than simple INIT grounding
  - Attentive MMT recovers more from impact of substantial masking

Future work

- Grounding as a way to reduce **biases** and improve robustness to **errors**

- Better models to balance **complementary** and **redundant** information

- Multimodality to resolve unknown words

  - The **dachshund** is running in the fields full of little white flowers.
  - O **UNK** corre no campo cheio de florzinhas brancas.
  - O cachorro corre no campo cheio de florzinhas brancas.

Minneapolis, MN

NAACL HLT 2019

June 2-7, 2019

Thank you!

References

Desmond **Elliott**, Stella Frank, Khalil Sima'an, and Lucia Specia. **2016**. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*. Association for Computational Linguistics, Berlin, Germany, pages 70–74.

Peter **Young**, Alice Lai, Micah Hodosh, and Julia Hockenmaier. **2014**. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Chiraag **Lala**, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. **2018**. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels, pages 630–637.

Desmond **Elliott**. **2018**. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 2974–2978.

Stig-Arne **Grönroos**, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. **2018**. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation*. Association for Computational Linguistics, Belgium, Brussels, pages 609–617.

References

Desmond **Elliott**, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. **2017**. Findings of the second shared task on multimodal machine translation and multilingual image description. In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Association for Computational Linguistics, Copenhagen, Denmark, pages 215–233.

Loïc **Barrault**, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. **2018**. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, pages 308–327.

Bryan A **Plummer**, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. **2015**. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In 2015 *IEEE International Conference on Computer Vision* (ICCV). pages 2641–2649.

Iacer **Calixto** and Qun Liu. **2017**. Incorporating global visual features into attention based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 992–1003.

Ozan **Caglayan**, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. **2017**. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark, pages 432–439.

References

Ozan **Caglayan**, Loïc Barrault, and Fethi Bougares. **2016**. Multimodal attention for neural machine translation. Computing Research Repository arXiv:1609.03976.

Iacer **Calixto**, Desmond Elliott, and Stella Frank. **2016**. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 634–638.

Jindřich **Libovický** and Jindřich **Helcl**. **2017**. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 196–202.

Jonathan H. **Clark**, Chris Dyer, Alon Lavie, and Noah A. Smith. **2011**. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 176–181.