

# Transform and Tell: Entity-Aware News Image Captioning

## Supplementary Material

Alasdair Tran, Alexander Mathews, Lexing Xie  
Australian National University

{`alasdair.tran`,`alex.mathews`,`lexing.xie`}@anu.edu.au

### 1. Live Demo

A live demo of our model is available at <https://transform-and-tell.ml>. In the demo, the user is able to provide the URL to a New York Times article. The server will then scrape the web page, extract the article and image, and feed them into our model to generate a caption.

### 2. Entity Distribution

Figure 1 shows how different name entity types are distributed in the training captions of the NYTimes800k dataset. The four most popular types are people’s names (PERSON), geopolitical entities (GPE), organizations (ORG), and dates (DATE). Out of these, people’s names comprise a third of all named entities. This motivates us to add a specialized face attention module to the model.

### 3. Model Complexity

Table 1 shows the number of training parameters in each of our model variants. We ensure that the total number of trainable parameters for each model is within 7% of one another (148 million to 159 million), with the exception of the model with face attention (171 million) and with object attention (200 million) since the latter two have extra multi-head attention modules.

### 4. Further Experimental Results

Table 2 reports BLEU-1, BLEU-2, BLEU-3, BLEU-4 [6] ROUGE [5], METEOR [2], and CIDEr [8]. Our results display a strong correlation between of all these metrics—a method that performs well on one metric tends to perform well on them all. Of particular interest is the CIDEr score since it uses Term Frequency Inverse Document Frequency (TF-IDF) to put more importance on less common words such as entity names. This makes CIDEr particularly well suited for evaluating news captions where uncommon words tend to be vitally important, e.g. people’s names.

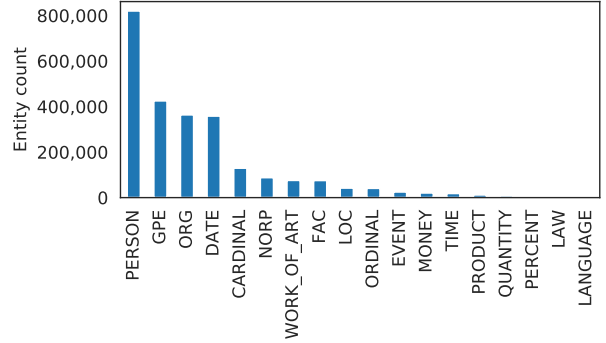


Figure 1: Entity distribution in NYTimes800k training captions. The four most common entity types are people’s names, geopolitical entities, organizations, and dates.

Table 3 further reports metrics on the entities. In particular, we show the precision and recall of all proper nouns and new proper nouns. We define a proper noun to be new if it has never appeared in any training caption or article text. This is in contrast to the rare proper noun metrics reported in the main paper, which are proper nouns that are not present in any training caption but might have appeared inside a training article context.

The three rightmost columns of Table 3 show the linguistic quality metrics, including caption length (CL), type-token ratio (TTR) [7], and Flesch readability ease (FRE) [3, 4]. The TTR is measured as

$$\text{TTR} = \frac{U}{W} \quad (1)$$

where  $U$  is the number of unique words and  $W$  is the total number of words in the caption. FRE is measured as

$$\text{FRE} = 206.835 - 1.015 \left( \frac{W}{S} \right) - 84.6 \left( \frac{B}{W} \right) \quad (2)$$

where  $W$  is the number of words,  $S$  is the number of sentences, and  $B$  is the number of syllables in the caption.

Table 1: Model complexity. See Table 3 caption in the main paper for more explanation of each model variant.

	No. of Parameters
LSTM + GloVe + IA	157M
Transformer + GloVe + IA	148M
LSTM + RoBERTa + IA	159M
Transformer + RoBERTa	125M
+ image attention (IA)	154M
+ weighted RoBERTa	154M
+ location-aware	154M
+ face attention	171M
+ object attention	200M

The higher TTR corresponds to a higher vocabulary variation in the text, while a higher FRE indicates that the text uses simpler words and thus is easier to read. Overall our models produce captions that are closer in length to the ground truths than the previous state of the art *Biten* [1]. Moreover, our captions exhibit a level of language complexity (as measured by Flesch score) that is closer to the ground truths. However, there is still a gap in TTR, Flesch, and length, between captions generated by our model and the human-written ground-truth captions.

Finally Figure 2 and Figure 3 show two further set of generated captions.

## References

- [1] Ali Furkan Biten, Lluís Gómez, Marçal Rusinol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [2] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.
- [3] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.
- [4] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.
- [5] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [7] Mildred C Templin. Certain language skills in children; their development and interrelationships. 1957.
- [8] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Table 2: BLEU, ROUGE, METEOR, and CIDEr metrics on the GoodNews and NYTimes800k datasets.

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
GoodNews	Biten (Avg + CtxIns) [1]	9.04	3.66	1.71	0.89	12.2	4.37	13.1
	Biten (TBB + AttIns) [1]	8.10	3.26	1.48	0.76	12.2	4.17	12.7
	LSTM + GloVe + IA	14.1	6.50	3.36	1.97	13.6	5.54	13.9
	Transformer + GloVe + IA	18.8	9.72	5.55	3.48	17.0	7.63	25.2
	LSTM + RoBERTa + IA	18.0	9.54	5.51	3.45	17.0	7.68	28.6
	Transformer + RoBERTa	19.7	11.3	6.96	4.60	18.6	8.82	40.9
	+ image attention	21.6	12.7	8.09	5.45	20.7	9.74	48.5
	+ weighted RoBERTa	22.3	13.4	8.72	6.0	21.2	10.1	53.1
	+ face attention	<b>22.4</b>	<b>13.5</b>	8.77	<b>6.05</b>	<b>21.4</b>	10.2	<b>54.3</b>
	+ object attention	<b>22.4</b>	<b>13.5</b>	<b>8.80</b>	<b>6.05</b>	<b>21.4</b>	<b>10.3</b>	53.8
NYTimes800k	LSTM + GloVe + IA	13.4	6.0	3.06	1.77	13.1	5.34	12.1
	Transformer + GloVe + IA	16.8	8.28	4.56	2.75	15.9	6.94	20.3
	LSTM + RoBERTa + IA	17.0	8.92	5.19	3.29	16.1	7.31	24.9
	Transformer + RoBERTa	18.2	10.2	6.37	4.26	17.3	8.14	33.9
	+ image attention	20.0	11.6	7.38	5.01	19.4	9.05	40.3
	+ weighted RoBERTa	20.9	12.5	8.18	5.75	19.9	9.56	45.1
	+ location-aware	21.8	13.5	8.96	6.36	21.4	10.3	52.8
	+ face attention	<b>21.6</b>	13.3	8.85	6.26	21.5	<b>10.3</b>	53.9
	+ object attention	<b>21.6</b>	<b>13.4</b>	<b>8.90</b>	<b>6.30</b>	<b>21.7</b>	<b>10.3</b>	<b>54.4</b>

Table 3: All proper noun and new proper noun precision (P) & recall (R) on the GoodNews and NYTimes800k datasets. Linguistic measures on the generated captions: caption length (CL), type-token ratio (TTR), and Flesch readability ease (FRE).

		All proper nouns		New proper nouns		CL	TTR	FRE
		P	R	P	R			
GoodNews	Ground truths	–	–	–	–	18.1	94.9	65.4
	Biten (Avg + CtxIns) [1]	16.5	12.2	2.70	12.0	9.89	92.2	78.3
	Biten (TBB + AttIns) [1]	19.2	11.0	4.21	12.3	9.14	90.7	77.6
	LSTM + GloVe + IA	16.1	11.3	0	0	14.0	89.5	77.2
	Transformer + GloVe + IA	22.7	18.4	0	0	16.0	88.4	73.9
	LSTM + RoBERTa + IA	25.1	20.8	1.68	7.86	15.0	89.0	75.7
	Transformer + RoBERTa	30.7	26.0	7.69	16.4	15.1	90.0	73.0
	+ image attention	33.4	28.0	8.53	19.3	15.2	90.0	72.5
	+ weighted RoBERTa	33.9	29.6	<b>15.2</b>	<b>24.4</b>	15.5	90.8	71.8
	+ face attention	34.3	29.8	13.6	22.2	15.4	90.8	71.8
	+ object attention	<b>34.7</b>	<b>29.9</b>	13.3	23.6	15.3	90.9	72.0
NYTimes800k	Ground truths	–	–	–	–	18.4	94.6	63.9
	LSTM + GloVe + IA	15.8	12.4	0	0	13.9	88.7	76.1
	Transformer + GloVe + IA	21.5	18.2	0	0	14.8	88.8	71.9
	LSTM + RoBERTa + IA	24.1	21.8	3.28	7.18	14.8	89.3	73.3
	Transformer + RoBERTa	28.0	26.0	13.4	14.5	15.2	90.4	71.4
	+ image attention	31.1	28.7	15.6	17.2	15.1	90.1	71.5
	+ weighted RoBERTa	31.8	30.5	21.7	20.2	15.5	91.6	70.1
	+ location-aware	36.4	34.1	26.3	<b>25.3</b>	15.1	91.7	70.8
	+ face attention	36.8	34.2	26.2	24.2	14.9	91.8	70.9
	+ object attention	<b>37.2</b>	<b>34.5</b>	<b>26.7</b>	25.1	14.8	91.9	71.2

Table 4: Geopolitical entity (GPE), organization (ORG), and date (DATE) precision (P) & recall (R) on the GoodNews and NYTimes800k datasets.

		GPE		ORG		DATE	
		P	R	P	R	P	R
GoodNews	Biten (Avg + CtxIns) [1]	12.0	11.5	5.67	7.45	6.12	4.03
	Biten (TBB + AttIns) [1]	12.8	8.41	5.81	7.36	5.86	4.06
	LSTM + GloVe + IA	15.6	12.8	14.0	8.58	11.0	8.20
	Transformer + GloVe + IA	20.8	18.8	16.6	11.8	12.0	10.1
	LSTM + RoBERTa + IA	20.8	19.2	16.9	12.3	13.4	10.9
	Transformer + RoBERTa	22.6	22.5	20.4	16.3	13.8	12.6
	+ image attention	<b>25.8</b>	24.5	21.0	17.3	14.4	13.0
	+ weighted RoBERTa	25.0	24.2	22.0	<b>18.7</b>	14.3	13.1
	+ face attention	24.9	24.4	21.6	18.5	14.7	<b>13.3</b>
	+ object attention	25.6	<b>24.7</b>	<b>22.4</b>	<b>18.7</b>	<b>15.1</b>	<b>13.3</b>
NYTimes800k	LSTM + GloVe + IA	16.0	14.7	8.60	4.89	11.3	8.31
	Transformer + GloVe + IA	19.1	21.8	12.1	7.95	11.3	10.1
	LSTM + RoBERTa + IA	20.2	22.2	13.1	8.95	11.8	11.1
	Transformer + RoBERTa	21.4	25.4	15.8	12.2	12.0	12.5
	+ image attention	23.9	27.3	17.6	13.6	12.8	13.2
	+ weighted RoBERTa	24.2	28.2	19.2	15.6	13.9	14.3
	+ location-aware	26.8	30.1	20.9	<b>17.3</b>	<b>14.1</b>	<b>14.1</b>
	+ face attention	<b>26.9</b>	<b>30.6</b>	20.7	16.5	13.9	<b>14.1</b>
	+ object attention	26.8	<b>30.6</b>	<b>21.9</b>	17.2	13.7	13.8


<p><b>An Artist Making a Powerful Statement — by Creating Work About Herself</b></p>  <p>During the final days of her solo show at Kravets Wehby Gallery in Manhattan this past spring, the mixed-media artist Theresa Chromati had something to confess about her latest body of work. “I realized that you can’t hide from anything,” she said, staring up at the 2019 painting “We All Look Back At It (Morning Ride).” In it, a nude figure squats with her glittering, butterfly-adorned buttocks in the air and her unobscured face turned to look directly at the viewer.</p> <p>For much of her career, Chromati, 26, depicted the naked bodies in her powerful portraits of black women behind protective disguises....</p>	<table> <tr> <td><b>Ground-truth caption</b></td><td>The mixed-media artist Theresa Chromati sits in front of an unfinished and currently untitled acrylic painting at her Brooklyn studio.</td></tr> <tr> <td><b>LSTM + GloVe + IA</b></td><td>“Untitled (Bubs),” 2017, oil on canvas.</td></tr> <tr> <td><b>Transformer + GloVe + IA</b></td><td>“Untitled (The Red Rose)” (2015), a painting by Nina Arianda.</td></tr> <tr> <td><b>LSTM + RoBERTa + IA</b></td><td>“The B-N-1,” by the artist and artist Ms. Chastain.</td></tr> <tr> <td><b>Transformer + RoBERTa</b></td><td>The artist Theresa Cromati in her studio in Manhattan.</td></tr> <tr> <td><b>+ image attention</b></td><td>The artist Theresa Cromati in her studio in New York.</td></tr> <tr> <td><b>+ weighted RoBERTa</b></td><td>“I’m a woman who’s not going to be a woman,” said the artist Theresa Cromati, who has been working with her own work since 2017.</td></tr> <tr> <td><b>+ location-aware</b></td><td>Theresa Nemati, who has created a new work, in her studio in Brooklyn.</td></tr> <tr> <td><b>+ face attention</b></td><td>Theresa Cromati in her studio in Brooklyn.</td></tr> <tr> <td><b>+ object attention</b></td><td>The artist Theresa Cromati in her studio in Manhattan.</td></tr> </table>	<b>Ground-truth caption</b>	The mixed-media artist Theresa Chromati sits in front of an unfinished and currently untitled acrylic painting at her Brooklyn studio.	<b>LSTM + GloVe + IA</b>	“Untitled (Bubs),” 2017, oil on canvas.	<b>Transformer + GloVe + IA</b>	“Untitled (The Red Rose)” (2015), a painting by Nina Arianda.	<b>LSTM + RoBERTa + IA</b>	“The B-N-1,” by the artist and artist Ms. Chastain.	<b>Transformer + RoBERTa</b>	The artist Theresa Cromati in her studio in Manhattan.	<b>+ image attention</b>	The artist Theresa Cromati in her studio in New York.	<b>+ weighted RoBERTa</b>	“I’m a woman who’s not going to be a woman,” said the artist Theresa Cromati, who has been working with her own work since 2017.	<b>+ location-aware</b>	Theresa Nemati, who has created a new work, in her studio in Brooklyn.	<b>+ face attention</b>	Theresa Cromati in her studio in Brooklyn.	<b>+ object attention</b>	The artist Theresa Cromati in her studio in Manhattan.
<b>Ground-truth caption</b>	The mixed-media artist Theresa Chromati sits in front of an unfinished and currently untitled acrylic painting at her Brooklyn studio.																				
<b>LSTM + GloVe + IA</b>	“Untitled (Bubs),” 2017, oil on canvas.																				
<b>Transformer + GloVe + IA</b>	“Untitled (The Red Rose)” (2015), a painting by Nina Arianda.																				
<b>LSTM + RoBERTa + IA</b>	“The B-N-1,” by the artist and artist Ms. Chastain.																				
<b>Transformer + RoBERTa</b>	The artist Theresa Cromati in her studio in Manhattan.																				
<b>+ image attention</b>	The artist Theresa Cromati in her studio in New York.																				
<b>+ weighted RoBERTa</b>	“I’m a woman who’s not going to be a woman,” said the artist Theresa Cromati, who has been working with her own work since 2017.																				
<b>+ location-aware</b>	Theresa Nemati, who has created a new work, in her studio in Brooklyn.																				
<b>+ face attention</b>	Theresa Cromati in her studio in Brooklyn.																				
<b>+ object attention</b>	The artist Theresa Cromati in her studio in Manhattan.																				

Figure 2: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The name “Chromati” has never appeared in the training data, and none of the models can spell the artist’s name correctly. They all miss the letter “h” in her name. Captions from models that use an LSTM or GloVe contain made-up names for both the painting and the artist. Finally the model that has no access to the image, *Transformer + RoBERTa*, still guesses correctly that the image is about the artist being in her studio. This shows that NYTimes article images can have a predictable theme.


<p><b>Bernie Sanders Heads to Canada for Affordable Insulin</b></p>  <p>WINDSOR, Ontario — Bernie Sanders wanted to make a point about a crippling injustice. So he crossed the border.</p> <p>Well, the northern one.</p> <p>On Sunday, he took about a dozen people with diabetes on a bus from Detroit to Windsor to get insulin at a Canadian pharmacy, just minutes from the border. Because of traffic, and multiple stops along the way, it took an hour and 17 minutes to get there and about the same time to get back. But the duration and the mileage were not really the main points.</p> <p>....</p>	<table> <tr> <td><b>Ground-truth caption</b></td><td>Mr. Sanders spoke for about four minutes outside the pharmacy.</td></tr> <tr> <td><b>LSTM + GloVe + IA</b></td><td>Senator Bernie Sanders of Vermont at a rally in the Bronx on Sunday.</td></tr> <tr> <td><b>Transformer + GloVe + IA</b></td><td>Senator Bernie Sanders of Vermont at a rally in Ottawa on Sunday.</td></tr> <tr> <td><b>LSTM + RoBERTa + IA</b></td><td>Senator Bernie Sanders of Vermont, center, with his wife, Ann, and son, Michael, at the Canadian border in Canada on Sunday.</td></tr> <tr> <td><b>Transformer + RoBERTa</b></td><td>Senator Bernie Sanders of Vermont, center, with a group of people with diabetes, at a Canadian pharmacy in Windsor, Ontario, on Sunday.</td></tr> <tr> <td><b>+ image attention</b></td><td>Senator Bernie Sanders of Vermont, who spoke to reporters on Sunday, took a group of people with diabetes to get a drug from Canada to Ottawa.</td></tr> <tr> <td><b>+ weighted RoBERTa</b></td><td>Senator Bernie Sanders of Vermont, center, at a rally in Detroit on Sunday.</td></tr> <tr> <td><b>+ location-aware</b></td><td>Mr. Sanders, center, and Ms. Nystrom, right, at a rally in Windsor, Ontario, on Wednesday.</td></tr> <tr> <td><b>+ face attention</b></td><td>Mr. Sanders spoke to reporters outside the pharmacy where he spoke about his son, Hunter, 22.</td></tr> <tr> <td><b>+ object attention</b></td><td>Mr. Sanders spoke to reporters outside the pharmacy in Windsor, Ontario, on Saturday.</td></tr> </table>	<b>Ground-truth caption</b>	Mr. Sanders spoke for about four minutes outside the pharmacy.	<b>LSTM + GloVe + IA</b>	Senator Bernie Sanders of Vermont at a rally in the Bronx on Sunday.	<b>Transformer + GloVe + IA</b>	Senator Bernie Sanders of Vermont at a rally in Ottawa on Sunday.	<b>LSTM + RoBERTa + IA</b>	Senator Bernie Sanders of Vermont, center, with his wife, Ann, and son, Michael, at the Canadian border in Canada on Sunday.	<b>Transformer + RoBERTa</b>	Senator Bernie Sanders of Vermont, center, with a group of people with diabetes, at a Canadian pharmacy in Windsor, Ontario, on Sunday.	<b>+ image attention</b>	Senator Bernie Sanders of Vermont, who spoke to reporters on Sunday, took a group of people with diabetes to get a drug from Canada to Ottawa.	<b>+ weighted RoBERTa</b>	Senator Bernie Sanders of Vermont, center, at a rally in Detroit on Sunday.	<b>+ location-aware</b>	Mr. Sanders, center, and Ms. Nystrom, right, at a rally in Windsor, Ontario, on Wednesday.	<b>+ face attention</b>	Mr. Sanders spoke to reporters outside the pharmacy where he spoke about his son, Hunter, 22.	<b>+ object attention</b>	Mr. Sanders spoke to reporters outside the pharmacy in Windsor, Ontario, on Saturday.
<b>Ground-truth caption</b>	Mr. Sanders spoke for about four minutes outside the pharmacy.																				
<b>LSTM + GloVe + IA</b>	Senator Bernie Sanders of Vermont at a rally in the Bronx on Sunday.																				
<b>Transformer + GloVe + IA</b>	Senator Bernie Sanders of Vermont at a rally in Ottawa on Sunday.																				
<b>LSTM + RoBERTa + IA</b>	Senator Bernie Sanders of Vermont, center, with his wife, Ann, and son, Michael, at the Canadian border in Canada on Sunday.																				
<b>Transformer + RoBERTa</b>	Senator Bernie Sanders of Vermont, center, with a group of people with diabetes, at a Canadian pharmacy in Windsor, Ontario, on Sunday.																				
<b>+ image attention</b>	Senator Bernie Sanders of Vermont, who spoke to reporters on Sunday, took a group of people with diabetes to get a drug from Canada to Ottawa.																				
<b>+ weighted RoBERTa</b>	Senator Bernie Sanders of Vermont, center, at a rally in Detroit on Sunday.																				
<b>+ location-aware</b>	Mr. Sanders, center, and Ms. Nystrom, right, at a rally in Windsor, Ontario, on Wednesday.																				
<b>+ face attention</b>	Mr. Sanders spoke to reporters outside the pharmacy where he spoke about his son, Hunter, 22.																				
<b>+ object attention</b>	Mr. Sanders spoke to reporters outside the pharmacy in Windsor, Ontario, on Saturday.																				

Figure 3: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The model that has no access to the image, *Transformer + RoBERTa*, is correct in predicting that the image is about Bernie Sanders. However it guesses that he is with a group of people with diabetes, which is not correct but is sensible given the article content. Some of the models manage to override the strong prior that he is at a rally (which is what many of Bernie Sanders images in the training set are about) and correctly say that he is outside a pharmacy. The caption from the model with object attention is the most accurate because it generates all three entities correctly: Windsor in Ontario, the reporters, and the pharmacy.