Incorporating Spatial Image Features into English Chinese Machine Translation

Nate Carlson

Project Objective

1. Generate a synthetic dataset for EN-CN Multimodal NMT
2. Build Multimodal Transformer model inspired by (Helcl et al., 2018)
3. Analyze ability of model to leverage visual input to improve translation

Expectations

- Visual information will have minimal impact on translation quality in cases where no noise is present in source text
- Multimodal model will perform better than baselines in cases where synthetic noise is introduced via masking
- Incorporating visual information into the model at decoding time will yield the best results

Related Work

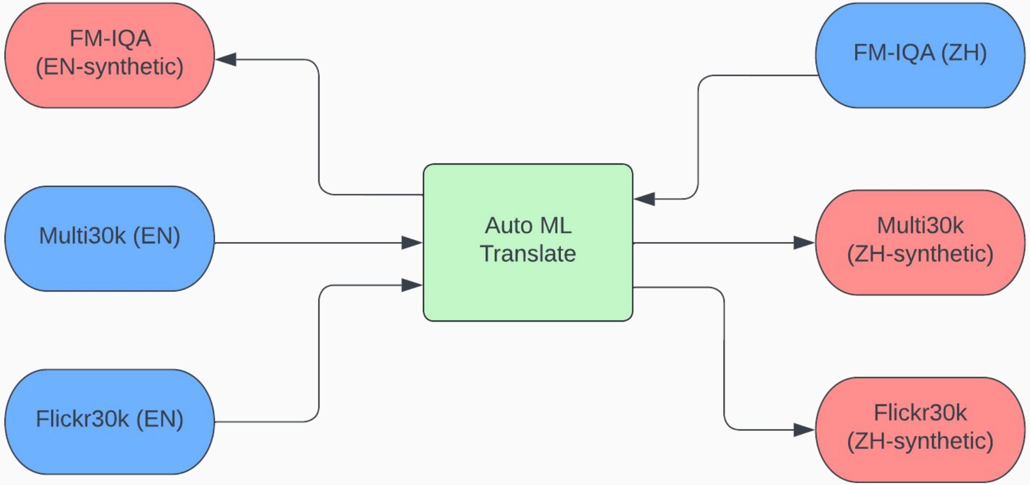
- RNN sequence models enhanced with global image features (Bahdanau et al., 2014)
- Attention between RNN hidden states and visual features (Caglayan et al., 2017)
- Transformer network replaces RNN with self-attention (Vaswani et al., 2017)
- Spatial image features in attention based transformer network (Helcl et al., 2018)

Project Objective

1. **Generate a synthetic dataset for EN-CN Multimodal NMT**
2. Build Multimodal Transformer model inspired by (Helcl et al., 2018)
3. Analyze ability of model to leverage visual input to improve translation

Dataset

- FM-QA
  - ~119,695 images
  - ~155,000 question answer pairs
- Flickr30k
  - ~30,000 images
  - ~150,000 descriptions
- Multi30k
  - Subset of flickr30k
  - ~30,000 images
  - ~30,000 descriptions
- Total
  - ~150,000 images
  - ~300,000 translation pairs

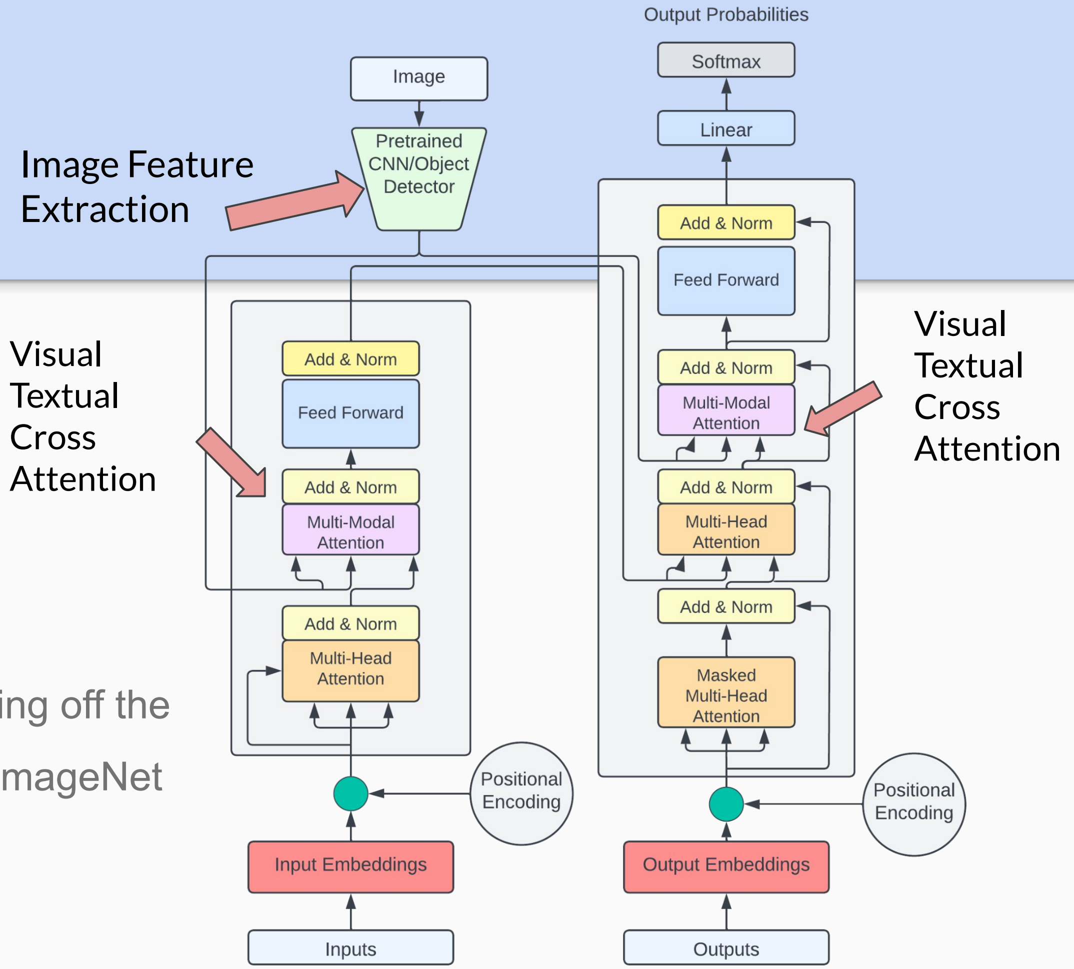


Project Objective

1. Generate a synthetic dataset for EN-CN Multimodal NMT
2. **Build Multimodal Transformer model inspired by (Helcl et al., 2018)**
3. Analyze ability of model to leverage visual input to improve translation

Model Architecture

- Doubly attentive transformer
  - Textual attention
  - Visual textual cross attention
- Image Feature Extraction
  - Image features are extracted using off the shelf ResNet-50 pre-trained on ImageNet (100,000+ classes)



Visual Textual Cross Attention

- **Queries**: output of first attention block in encoder/decoder
- **Keys**, **Values**: projected image feature maps
- Allows model to add visual information to intermediary context vectors

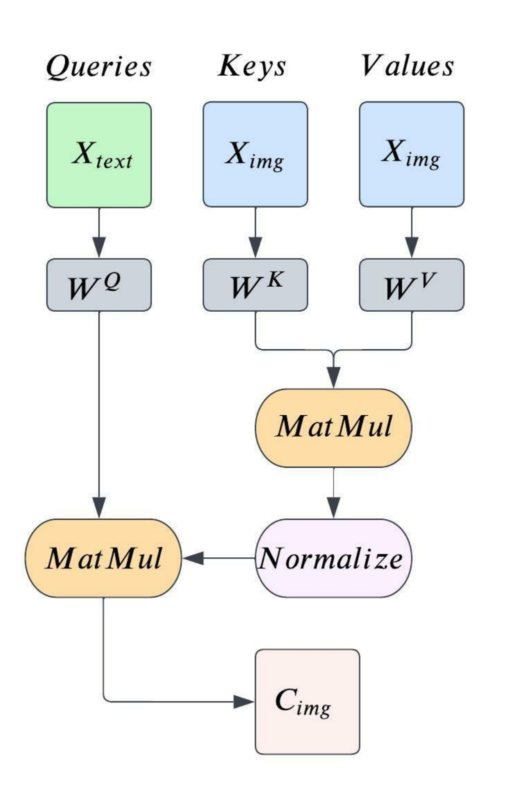
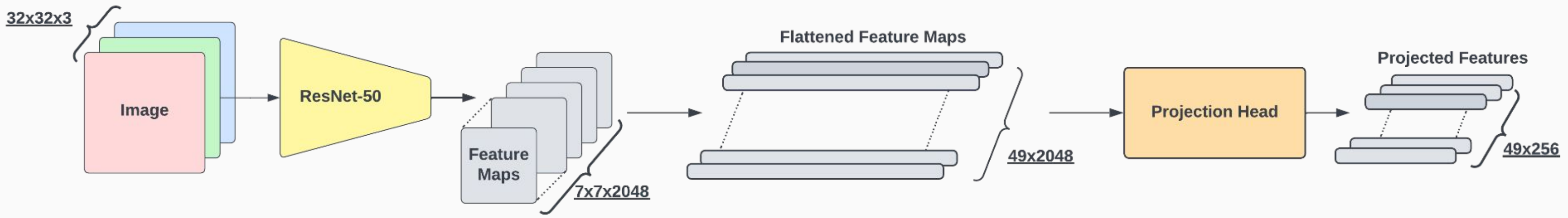


Image Feature Extraction

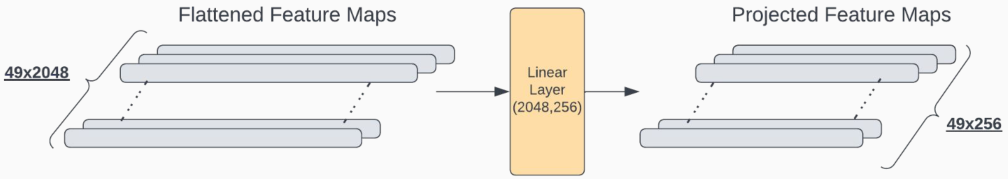
- 3 Steps:
  - **Extract**
  - **Flatten**
  - **Project**



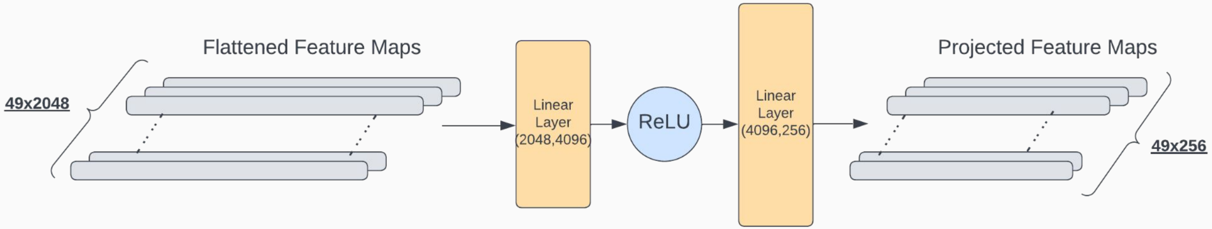
Projection Head

- Linear projection head + dropout(0.1)
- Nonlinear projection head (ReLU) + dropout(0.1)

Linear Projection Head



Nonlinear Projection Head



Training Setup

- Training
  - Hyper Parameters
    - 20 epochs
    - 6 encoder layers
    - 6 decoder layers
    - 8 attention heads
    - hidden dimension 256
    - Dropout probability 0.1
  - Noam Optimizer (2000 warm up steps, betas 0.9 & 0.98)
  - Label Smoothing

- Data: Only used Flickr30k + Multi30k due to compute constraints
- 80%-10%-10% Train-Validation-Test split
  - Train: 55,127 instances
  - Validation: 6,891 instances
  - Test: 6,891 instances

Tools & Programming Languages

- Coded all in Pytorch
- For my base model I used code from the blog post below. I adapted the same code for the multimodal models.
  - https://cuicaihao.com/the-annotated-transformer-english-to-chinese-translator/

Project Objective

1. Generate a synthetic dataset for EN-CN Multimodal NMT
2. Build Multimodal Transformer model inspired by (Helcl et al., 2018)
3. **Analyze ability of model to leverage visual input to improve translation**

Evaluation Metrics

- BLEU: $$BLEU = BP * exp(\sum_{k=1}^{n} w_k log(p_k))$$
- chrF+: $$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}$$

Experiments

- Experiment 1: Where and how to incorporate visual context?
  - Encoder
  - Decoder
  - Encoder + Decoder
  - Image Projector
    - Linear or Nonlinear?
- Experiment 2: Ablation Studies
  - Blank images
  - Random images
- Experiment 3: Source Degradation
  - **Probabilistic POS Masking**: Mask nouns, verbs, adj w/ probability 0.3
  - **Deterministic Masking**: mask 2nd half of source sentence
- Experiment 4: Human Evaluations
  - 200 sentences evaluated in DQF framework

Results

Experiment 1: Where and how to incorporate visual context?

	BLEU	chrF
base	54.27	48.31
enc + linear proj	53.57	46.93
enc + nonlinear proj	**54.73***	48.23
dec + linear proj	54.13	47.45
dec + nonlinear proj	**55.07*~**	**48.99*~**
enc + dec + linear proj	49.77	43.75
enc + dec + nonlinear proj	**54.50***	**48.53***

*** : outperformed baseline**
~: best performing model

Expectations

- **Visual information will have minimal impact on translation quality in cases where no noise is present in source text**
- **Incorporating visual information into the model at decoding time will yield the best results**
- Multimodal model will perform better than baselines in cases where synthetic noise is introduced via masking

Results

Experiment 2: Ablation Studies

(dec + nonlinear proj)	BLEU	chrF
true images	**55.07~**	**48.99~**
blank images	54.81	48.57
random images	54.14	48.79

~: best performing model

Results

Experiment 3: Source Degradation

Probabilistic POS Masking		
	BLEU	**chrF**
base	38.01	32.59
dec+nonlinear proj	**39.37~**	**33.74~**

~: best performing model

Deterministic Masking		
	BLEU	**chrF**
base	28.48	25.13
dec+nonlinear proj	**30.66~**	**27.03~**

Expectations

- Visual information will have minimal impact on translation quality in cases where no noise is present in source text
- Incorporating visual information into the model at decoding time will yield the best results
- **Multimodal model will perform better than baselines in cases where synthetic noise is introduced via masking**

Sample Outputs: Probabilistic POS Masking

src: people are in a laundry mat washing clothes .

masked src: people are in a laundry BLANK washing clothes .

base: 人们在黑色垫子上穿着深色衣服。(People wear dark clothing on black mats.)

multimodal: 人们在洗衣垫上工作。(People are working in a laundry mat)

src: a man gives a fish to a boy .

masked src: a man gives a BLANK to a boy .

base: 一个人指着一个女人的手。(A man points to a woman's hand.)

multimodal: 一个男人把一个鱼线递给一个男人。(A man hands a fishing line to a man)

Sample Outputs: Probabilistic POS Masking

src: a family is posing with spongebob squarepants .

masked src: a BLANK is posing with BLANK squarepants .

base: 一个女孩正在摆姿势。(a girl is posing)

multimodal: 一个男人正在和海绵宝宝合影。(a man is taking a photo with spongebob squarepants)

src: man blows bubbles in a bathtub .

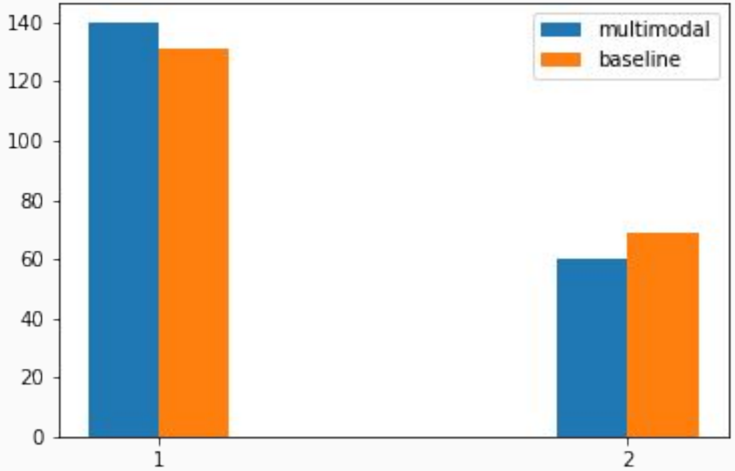
masked src: man blows BLANK in a bathtub .

base: 男子在田野里打篮球。(man playing basketball in field)

multimodal: 男人在浴缸里吹泡泡。(a man blows bubbles in a bathtub)

Results

Experiment 4: Human Evaluations



Conclusions

- Visual information is most useful at when incorporated at decoding time
- Adding a nonlinear projection head to the images improves translation
- Visual information is most useful when there are gaps in source text
- Having the correct image bitext pairs yields the best results

Future Work

- Post editing of synthetic dataset generated in this project
- Training a larger model with all of the data
- Exploring other architectures

References

- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv. https://doi.org/10.48550/ARXIV.1409.0473
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. arXiv. https://doi.org/10.48550/ARXIV.1706.03762
- Helcl, J., Libovický, J., & Variš, D. (2018). CUNI System for the WMT18 Multimodal Translation Task. Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 616–623. https://doi.org/10.18653/v1/W18-6441
- Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., & van de Weijer, J. (2017). LIUM-CVC Submissions for WMT17 Multimodal Translation Task. CoRR, abs/1707.04481. http://arxiv.org/abs/1707.04481
- https://github.com/cuicaihao/Annotated-Transformer-English-to-Chinese-Translator