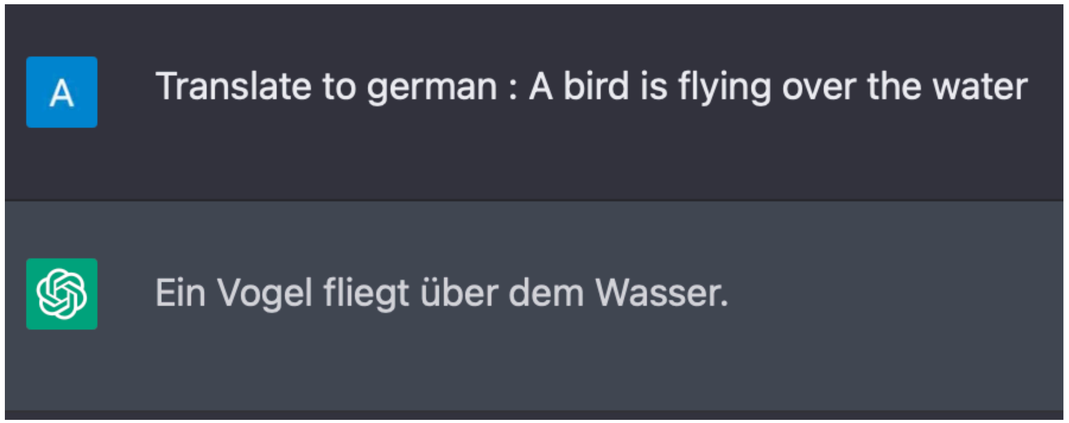
VALHALLA: VisuAL HALLucinAtion for Machine Translation

Li et al. 2022

Presenter: Alvi Ishmam

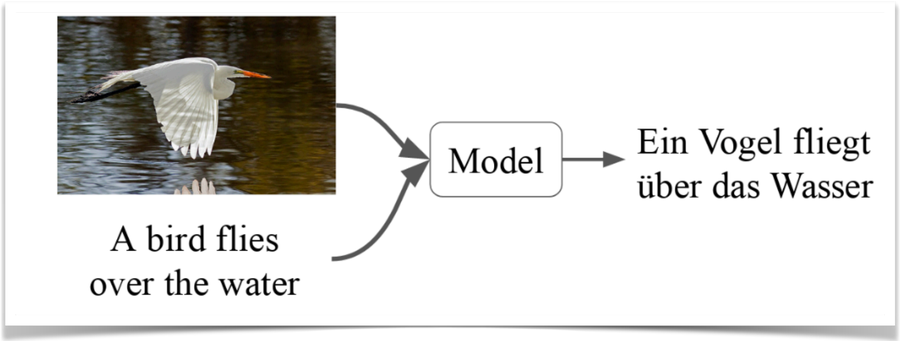
CS 6804, 03/15/2023

Machine Translation



Use of computer systems that translates text or speech from one language to another

Multimodal Machine Translation



Input: text, image, audio..
Output: target language text

Machine Translation (MT) Systems

- Rule based (Nyberg et al. 1992)
  - Knowledge based parsing and mapping form source to target language
- Statistical MT approaches (Lopez et al. 2008, Koehn 2007)
  - Mapping of source to target language done by confusion network - a decoder
- End to end neural networks (Bandanau et al. 2014, Sustskever et al. 2014)
  - Soft search of part of source sentence that are relevant to target sentence

However....

- These approaches only address text to text translation
- Other modalities provide useful rich semantic information

Learning Side Information with Modality Hallucination

Hoffman et al. 2016

Overview

- RGB object detection model
- Multiple modalities produce better recognition model
- RGB image capturing devices are more pervasive than depth capturing devices
- The recognition model need to perform well on RGB images alone as input

Learning Side Information with Modality Hallucination

Contribution

- Incorporate additional side information (**Depth information**) as an **image modality** at training time
- A representation learning through hallucination network
- Final model only sees RGB images in test times
- Performs better than single modality in NYUD2 dataset

Related Works

- **RGB-D Detection**
  - Adding additional depth network representation into CNN (Gupta et al. 2015)
- **Transfer Learning**
  - Mapping missing modalities through hallucination (Christoudias et al. 2010)
- **Learning with Side Information**
  - Additional data/ privileged information can be utilized in bonding box, image tags (Sharmanska et al 2013)

Methodology: Overview

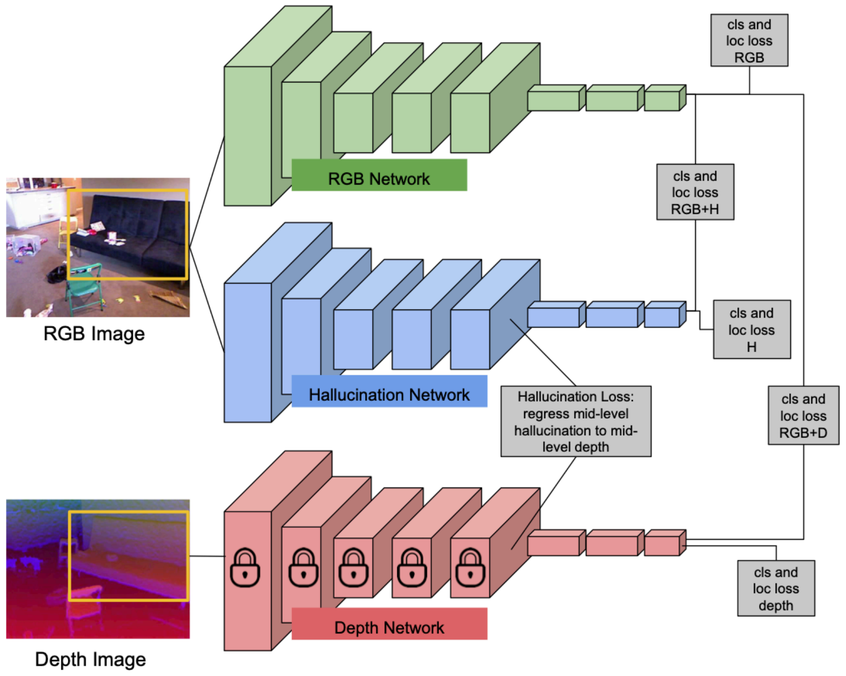


Figure 1: Training our modality hallucination architecture. We learn a multimodal Fast R-CNN [10] convolutional network for object detection. Our hallucination branch is trained to take an RGB input image and mimic the depth mid-level activations. The whole architecture is jointly trained with the bounding box labels and the standard softmax cross-entropy loss.

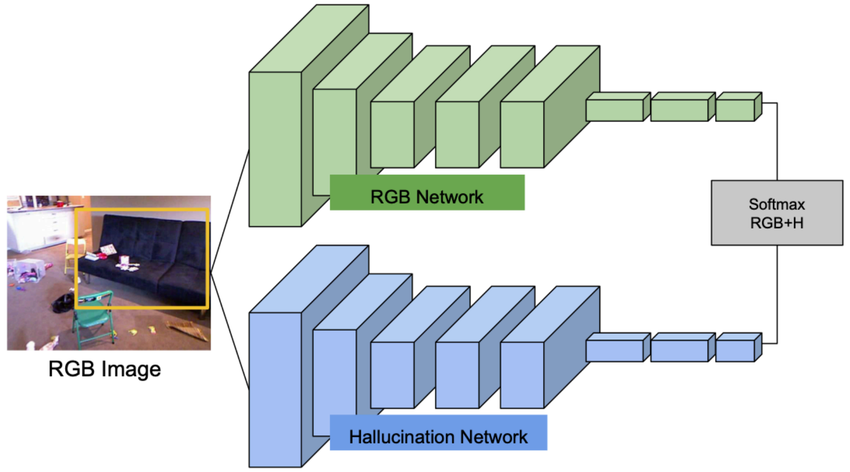


Figure 2: Test time modality hallucination architecture.

- Base model: Multilayer CNN for all networks
- Two modalities: 1) RGB 2) privileged depth modality with ***Hallucination*** network
- Hallucination net:
  - Inputs: RGB, image regions
  - Output: Detection scores for each category for each region

Methodology

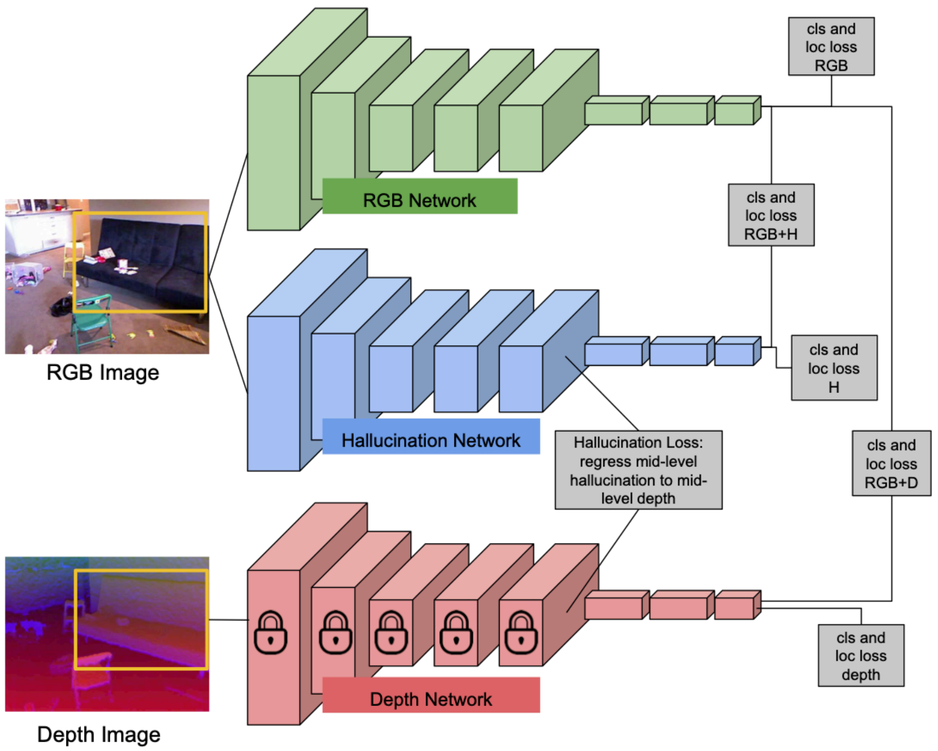


Figure 1: Training our modality hallucination architecture. We learn a multimodal Fast R-CNN [10] convolutional network for object detection. Our hallucination branch is trained to take an RGB input image and mimic the depth mid-level activations. The whole architecture is jointly trained with the bounding box labels and the standard softmax cross-entropy loss.

Hallucination loss: $\mathcal{L}_{\text{hallucinate}}(\ell) = \|\sigma(A_\ell^{\text{dNet}}) - \sigma(A_\ell^{\text{hNet}})\|_2^2$

Final Optimization

- 5 softmax cross entropy losses using bonding box labels as target, L_{cls}
- 5 smooth l1 loss using bonding box co-ordinates as targets, L_{loc}
- 1 additional hallucination loss

$$\begin{aligned}\mathcal{L} = \; & \gamma \mathcal{L}_{\text{hallucinate}} \\ & + \alpha \left[\mathcal{L}_{\text{loc}}^{\text{dNet}} + \mathcal{L}_{\text{loc}}^{\text{rNet}} + \mathcal{L}_{\text{loc}}^{\text{hNet}} + \mathcal{L}_{\text{loc}}^{\text{rdNet}} + \mathcal{L}_{\text{loc}}^{\text{rhNet}}\right] \\ & + \beta \left[\mathcal{L}_{\text{cls}}^{\text{dNet}} + \mathcal{L}_{\text{cls}}^{\text{rNet}} + \mathcal{L}_{\text{cls}}^{\text{hNet}} + \mathcal{L}_{\text{cls}}^{\text{rdNet}} + \mathcal{L}_{\text{cls}}^{\text{rhNet}}\right]\end{aligned}$$

Experiments

Datasets/ Base Network Setup

- RGB-D detection dataset NYUD2 (Ye 2013)
  - 1449 labeled image
- Base Network:
  - Fast R-CNN modified to AlexNet or VGG-1024 for all modules (RGB, depth, hallucination)
  - Depth images are encoded using HHA encoding
  - Depth network is fine-tuned after initializing with RGB weights
- Hyperparameters:
  - Update all layers with same learning rate

Experiments

NYUD2 Detection Evaluation

method	btub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB only [10] (A)	7.5	50.6	**36.8**	1.4	30.2	34.9	10.8	21.5	27.8	16.9	26.0	32.6	20.6	**25.1**	31.6	**36.7**	14.8	25.1	54.6	26.6
RGB ensemble (A-A)	10.5	53.7	33.6	1.6	32.0	34.8	12.2	20.8	34.5	19.6	28.6	45.7	28.5	24.4	31.4	34.7	14.5	**34.0**	**56.1**	29.0
Our Net (A-RGB, A-H)	**13.9**	**56.1**	34.4	**1.9**	**32.9**	**40.5**	**12.9**	**22.6**	**37.4**	**22.0**	**28.9**	**46.2**	**31.9**	22.9	**34.2**	34.2	**19.4**	33.2	53.6	**30.5**
RGB only [10] (V)	15.6	59.4	38.2	1.9	33.8	36.3	12.1	24.5	31.6	18.6	25.5	46.5	30.1	20.6	30.3	40.5	19.5	**37.8**	45.7	29.9
RGB ensemble (A-V)	14.8	60.4	**43.1**	**2.1**	36.4	40.7	13.3	**27.1**	35.5	20.8	29.9	**52.9**	**33.5**	26.2	33.0	44.4	19.9	36.7	50.2	32.7
Our Net (A-RGB, V-H)	**16.8**	**62.3**	41.8	**2.1**	**37.3**	**43.4**	**15.4**	24.4	**39.1**	**22.4**	**30.3**	46.6	30.9	**27.0**	**42.9**	**46.2**	**22.2**	34.1	**60.4**	**34.0**

Table 1: **Detection (AP%) on NYUD2 test set:** We compare our performance (pool5 hallucinate) against a Fast R-CNN [10] RGB detector trained on NYUD2 and against an ensemble of Fast R-CNN RGB detectors. AlexNet [21] architecture is denoted as 'A' and VGG-1024 [10, 28] architecture is denoted as 'V'. Our method outperforms both the RGB-only baselines and the RGB ensemble baselines.

Experiments

Hallucination net initialization

Initial Weights	bathtub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB	7.5	50.4	9.9	**0.9**	**26.2**	**24.9**	5.8	**15.8**	13.0	**29.8**	12.0	43.1	20.9	14.7	17.9	25.3	**15.1**	32.5	59.1	22.4
depth	9.9	**52.4**	**14.9**	**0.9**	24.9	24.4	4.3	15.3	18.1	24.1	**14.8**	**45.8**	**27.2**	**18.5**	**21.3**	**29.0**	13.7	**33.6**	**66.4**	**24.2**
random	**10.5**	47.6	12.3	0.6	23.5	20.2	**6.0**	13.0	**19.3**	12.0	13.3	42.8	12.8	12.1	13.6	23.0	13.9	28.6	61.5	20.3

Table 2: **RGB Detection (AP%) on NYUD2 val set:** We compare intializing the hallucination network by randomly initializing or by using the pre-trained RGB or depth parameter values.

Experiments

Which layer to Hallucinate

hallucination layer	bathtub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB only	4.9	45.5	10.9	1.3	21.5	23.6	5.4	14.5	12.7	17.4	9.4	40.9	17.2	14.9	19.9	19.2	14.0	32.5	66.3	20.6
pool1	**12.0**	**54.0**	**17.9**	**1.1**	24.5	23.6	5.0	15.2	16.3	12.7	13.3	40.0	24.7	16.6	20.5	**29.6**	14.9	27.4	55.3	22.3
pool2	8.4	50.7	13.5	1.0	24.2	26.0	6.6	13.1	13.8	17.8	11.7	40.7	21.8	15.0	20.5	22.4	15.2	27.2	59.7	21.5
conv3	8.8	52.5	13.2	1.0	25.6	26.2	3.3	13.2	14.9	17.0	**16.2**	41.6	22.2	**20.2**	22.9	24.6	**17.2**	**37.4**	65.6	23.3
conv4	9.7	51.2	12.9	1.0	26.3	**26.8**	**6.9**	**17.4**	16.7	22.0	12.4	43.2	15.5	16.4	**24.0**	23.5	16.2	34.2	64.2	23.2
pool5	9.9	52.4	14.9	0.9	24.9	24.4	4.3	15.3	**18.1**	**24.1**	14.8	**45.8**	**27.2**	18.5	21.3	29.0	13.7	33.6	**66.4**	**24.2**
fc6	10.3	47.2	12.0	0.6	21.7	20.0	5.9	12.8	13.8	20.5	11.8	34.4	16.3	13.1	14.8	27.3	16.1	28.8	60.5	20.4
fc7	3.3	49.4	12.7	0.8	24.1	21.8	4.8	15.2	16.8	11.7	10.0	43.4	18.7	14.2	20.6	25.2	14.4	29.5	63.1	21.0
fc8	4.2	50.7	13.9	0.9	23.8	23.6	5.4	15.5	18.0	13.2	13.3	42.0	20.9	15.8	22.3	23.8	14.5	29.6	63.6	21.8

Table 3: **RGB Detection (AP%) on NYUD2 val set:** We compare hallucinating different mid-level features with our method.

Takeaway

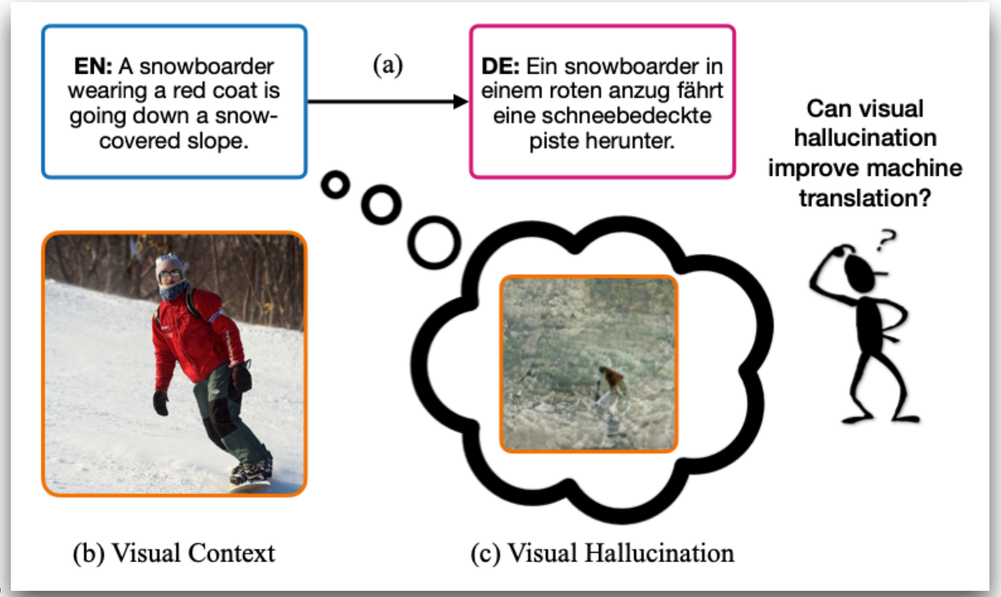
Hallucination networks improves object detection performance by adding additional side information to the model

VALHALLA: VisuAL HALLucinAtion for Machine Translation

Li et al. 2022

Motivation

- Machine Translation system relies on ***text*** only
- Lack explicit grounding from other modalities
- Additional visual context should help MT
- Grounding should improve data efficiency
- Benefit translation process in low resource settings



Related Works

- **Multimodal Machine Translation**
  - ✦ Visual alignment for unsupervised word mapping (Sigurdsson et al. 2020)
  - ✦ Beneficial under limited textual context (Caglayan 2019, Elliot et al. 2018)
- **Vision-Language Leaning**
  - ✦ Visual grounding is used for MT in many tasks e.g., semantic parsing, coreference resolution, representation learning (Shi et al. 2019)
  - ✦ Still improving MT with no visual information in test time remains as an open challenge
- **Modality Hallucination**
  - Similar concept is applied in object detection task where depth features are hallucinated from RGB input (Hoffman 2016)

Methodology

Traditional Machine Translation (Text to Text)

- Input source sentence $x = (x_1, \ldots, x_S)$,
- Output target sentence $y = (y_1, \ldots, y_T)$
- Transformer, $f_T = (f_T^{enc}, f_T^{dec})$
- The likelihood of target tokens conditioned on input sentence
- Transformer is trained optimizing cross entropy loss

$$p\left(y \mid x; \mathbf{f_T}\right) = \prod_{i=1}^{T} \mathbf{f_T}\left(y_i \mid y_{<i}, x\right)$$
$$\triangleq \prod_{i=1}^{T} \mathbf{f_T^{dec}}\left(y_i \mid y_{<i}, \mathbf{f_T^{enc}}(x)\right),$$

Methodology

Multimodal Machine Translation (Text + Visual to Text)

- Input source sentence $x = (x_1, \ldots, x_S)$,
- Input visual information: v
- Output target sentence $y = (y_1, \ldots, y_T)$
- Transformer, $f_T = (f_T^{enc}, f_T^{dec})$
- Use encoder, f_v to map into latent visual representation $z = f_v(v)$
- Transformer is trained optimizing cross entropy loss
- Issues: requires both text and image during inference

$$p\left(y \mid x; \mathbf{f_T}\right) = \prod_{i=1}^{T} \mathbf{f_T}\left(y_i \mid y_{<i}, x\right)$$

$$\triangleq \prod_{i=1}^{T} \mathbf{f_T^{dec}}\left(y_i \mid y_{<i}, \mathbf{f_T^{enc}}(x)\right),$$

Text to Text

$$p\left(y \mid x, z; \mathbf{f_T}\right) = \prod_{i=1}^{T} \mathbf{f_T}\left(y_i \mid y_{<i}, x, z\right).$$

Text + Visual to Text

Methodology: VALHALLA Architecture

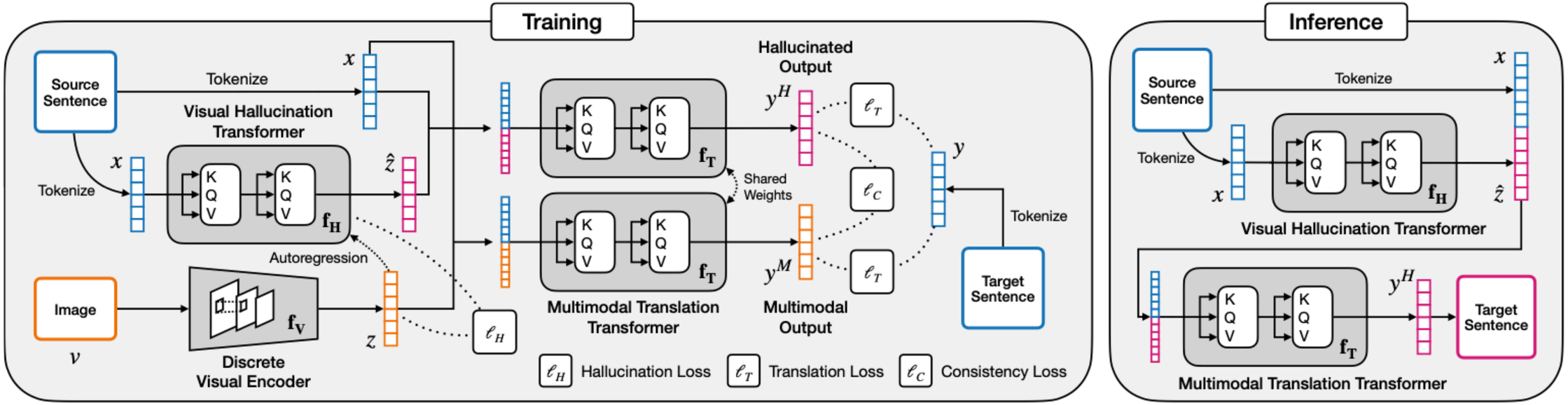


Figure 2: Overview of **VALHALLA** Architecture for Machine Translation. **Left**: Training pipeline of **VALHALLA**. Translation outputs are gathered from two streams of input, either with ground-truth visual tokens z or hallucinated representation $\hat{z}$, and optimized on a combination of *hallucination*, *translation* and *consistency* losses. **Right**: Inference process of **VALHALLA** in the absence of visual inputs.

Methodology: VALLHALLA Architecture

Choice of Visual Encoder: Continuous vs Discrete

- Traditionally combine text input with ***continuous*** visual embeddings extracted from pre-trained models e.g., ResNet50, ViT
- Issues:
  - Difficult to concatenate with discrete word embeddings with continuous visual embeddings
  - Complex aggregation module is required
  - Training instability
  - Model prediction collapse to mean value

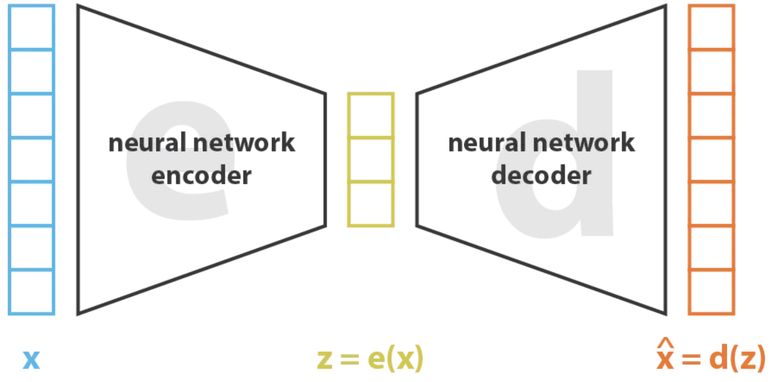
Idea:

Combine discrete text embeddings with ***discrete*** visual encoder

- Easy to aggregate between the modalities
- Easier to optimize with vanilla cross entropy loss

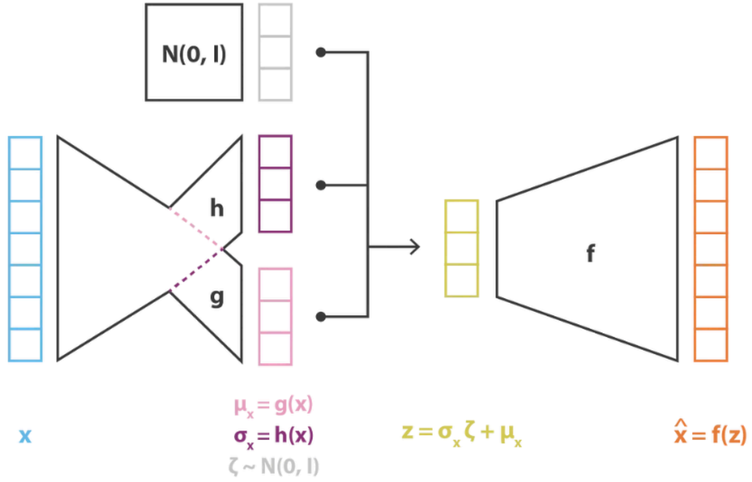
Methodology: VALLHALLA Architecture

Discrete Visual Encoder: Autoencoder



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

Auto Encoder



$$\text{loss} = C \| x - \hat{x} \|^2 + KL[\, N(\mu_x, \sigma_x), N(0, I) \,] = C \| x - f(z) \|^2 + KL[\, N(g(x), h(x)), N(0, I) \,]$$

Variational Autoencoders representation.

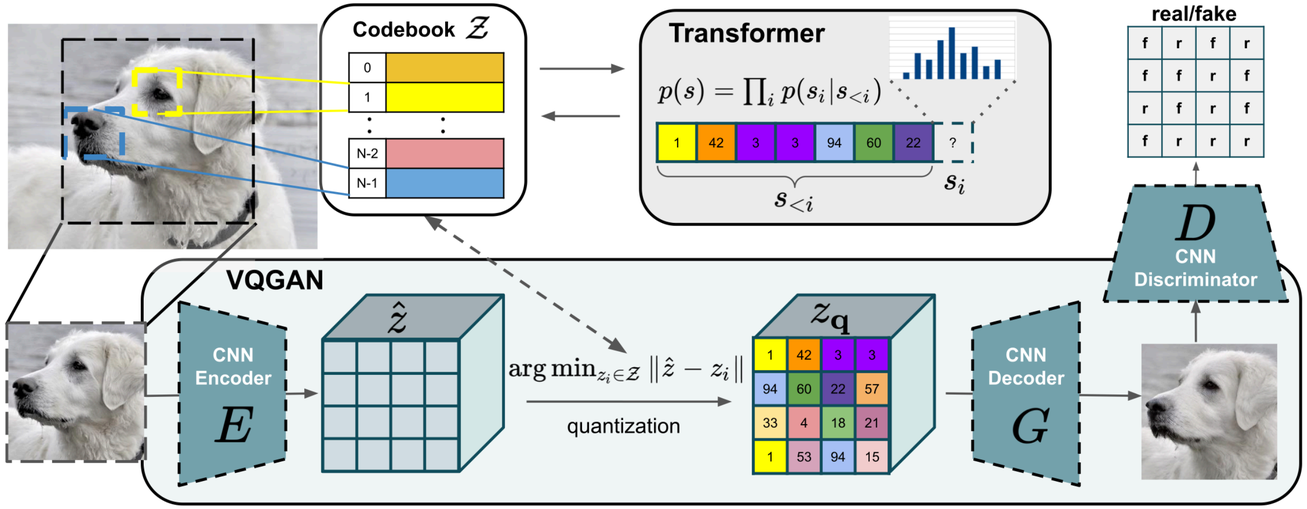
- Due to overfitting, the latent space can be very irregular
- Close points in latent space can give different decoded data
- Is not reliable for generating new data

- Distribution over the latent space instead a single point
- Add a regularization term to loss function for better organization

Image source: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Methodology: VALLHALLA Architecture

Variational Quantization GAN with VAE : VQGAN VAE (ESSER et al. 2021)



- Learn to represent image with codes from a learned, discrete codebook, Z
- A codebook provides an interface between transformers
- Obtain the representation from the encoder and subsequent element-wise quantization onto its closest codebook entry

- High resolution image requires globally consistent and locally realistic pattern generation from the model
- Pixel level representation is not sufficient
- Representation of rich semantic information in a form of codebook
- Codebook constituents example: Bounding box of mouth, nose, ear, rather than a full face

Methodology: VALLHALLA Architecture

Discrete Visual Encoder based on VQGAN VAE

- Implement vector quantization with VQGAN VAE

$$z = \mathcal{Q}(\mathbf{f_V}(v); E_V).$$

- Q is quantization function and Ev is d dimensional visual codebook of size k

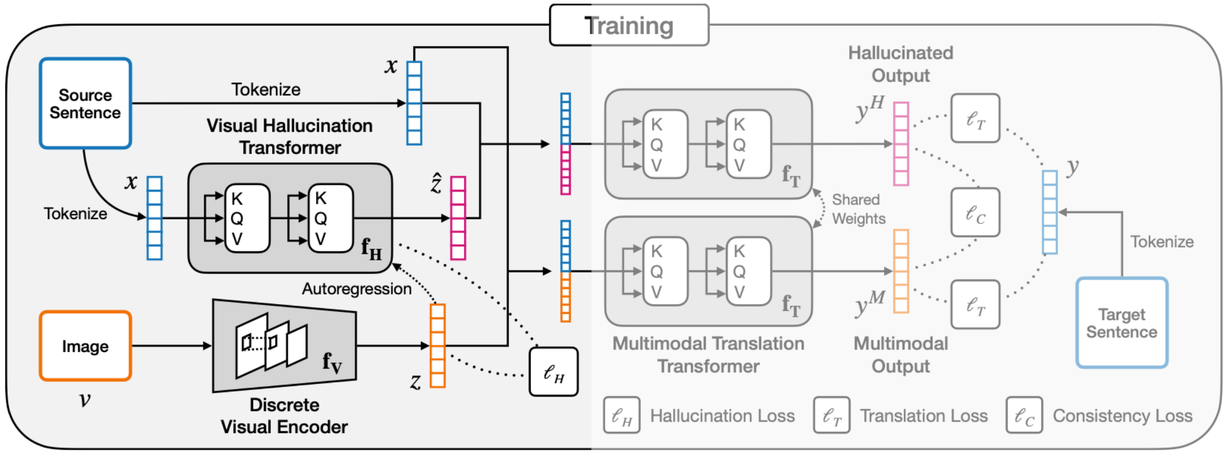
$$\mathcal{Q}_i(c; E_V) = \underset{k \in \{1,\ldots,K\}}{\arg\min} \|c_i - e^{(k)}\|_2$$

- Maps each spatial location of feature array c to the closest index of codebook e

Methodology: VALLHALLA Architecture

Visual Hallucination:

- VALHALLA relies on this module when no visual information is available during testing
- Uses an autoregressive transformer (Ramesh et al. 2021)
- Autoregressively model text and image tokens as a single stream of data
- Maximize the joint likelihood of x and z by optimizing the cross-entropy hallucination loss
- Final hallucinated visual sequences is defined by the most likely tokens predicted by f_h at each timestamp



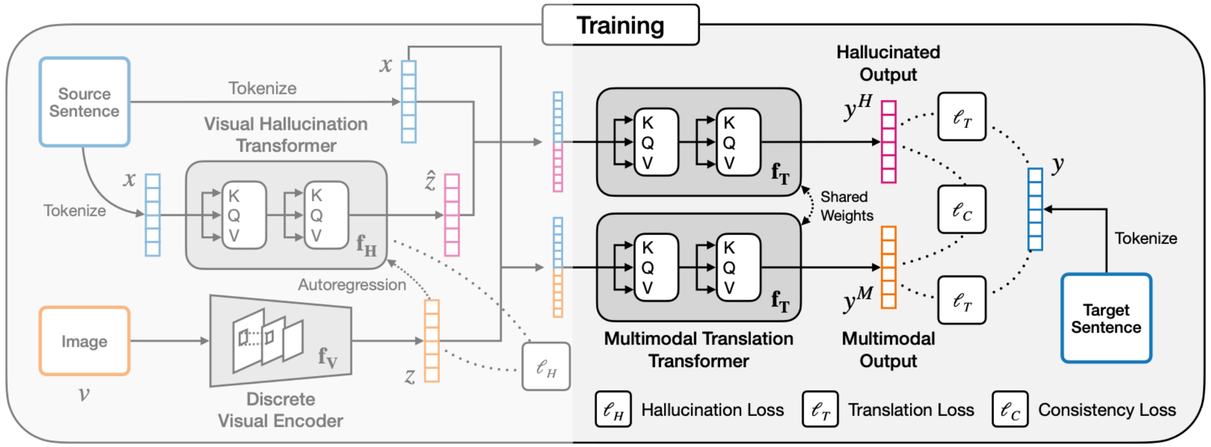
$$\ell_H(\mathbf{f_H}) = \mathbb{E}_{(x,z)}\left[-\log p(x, z; \mathbf{f_H})\right].$$

$$p(x, z; \mathbf{f_H}) = p(x; \mathbf{f_H}) p(z \mid x; \mathbf{f_H})$$
$$= \prod_{i=1}^{S} \mathbf{f_H}(x_i \mid x_{<i}) \prod_{j=1}^{V} \mathbf{f_H}(z_j \mid z_{<j}, x).$$

$$\hat{z}_i = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \mathbf{f_H}(z_i = k \mid z_{<i}, x),$$

Methodology: VALLHALLA Architecture

Visual Hallucination cont



- Issue: Creates mismatch between training and testing when decoding is performed only on source text tokens
- Constancy loss is computed:

$$\ell_C(\mathbf{f_H}, \mathbf{f_T}) = \mathbb{E}_{(x,z,y)} \left[\sum_{i=1}^{T} \mathrm{KL}[y_i^M \parallel y_i^H] \right]$$

- y_i^M, y_iH are the next word distributions from ground-truth visual tokens and hallucinated features respectively,
- $\mathrm{KL}[y_i^M || y_iH]$ is the Kullback-Leibler divergence between the two conditional distributions.

Methodology: VALLHALLA Architecture

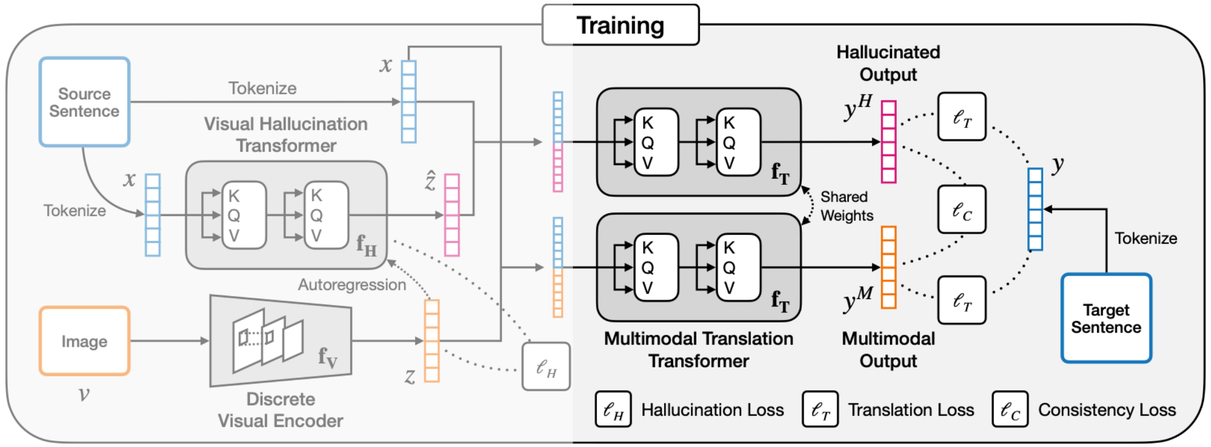
Visual Hallucination cont: Optimization

- Joint optimization of consistency loss and translation loss is critical
- The output of the hallucination transformer ($\hat{z}$) prevents loss gradients from back propagating.
- **Idea**: Relax the training objective rather than taking the most likely predicted tokens from the transformer
- Gumbel-softmax relaxation (Maddison 2016) during training,

$$\hat{z}_i = \sum_{k=1}^{K} \frac{\exp((\log \pi_{i,k} + g_k)/\tau)}{\sum_l \exp((\log \pi_{i,l} + g_l)/\tau)} o_k$$

Final optimization objective of VALHALLA:

$$\ell(\mathbf{f_H}, \mathbf{f_T}) = \ell_T(\mathbf{f_T}; z) + \ell_T(\mathbf{f_T}; \hat{z}) + \gamma_H \ell_H(\mathbf{f_H}) + \lambda_C \ell_C(\mathbf{f_H}, \mathbf{f_T}),$$



τ is the temperature of the softmax

o_k is the hot vector of length k, activated at dimension k, g1,g2, ... ~Gumbel (0,1)

Experiments

Datasets and Tasks

- **Datasets - Multi30K (Elliot 2016), Wikipedia Image Text (WIT) (Srinivasan et al. 2021) and WMT2014 (Bojar 2014)**
- Three test split; **Test2016, Test2017, MSCOCO**
- WIT is created by extracting text-image pairs from Wikipedia
- New benchmarks using three different settings
- **Tasks**: **Well resourced** (EN -> {DE, FR}), **Under resourced** ({EN -> RO}), **Non English** (DE-> ES)
- WMT does not provide aligned image
- Used CLIP (Radford et al. 2021) to retrieve top-5 images from Multi30K or WIT datasets to train transformers.

.

Experiments

Models, Baselines

- **Models**: Three variants; Base, Small, Tiny
- Multi30K : Small and Tiny due to better performance
- WMT/ WIT: Base (Well resourced), small (under/non English task)
- **Baselines:**
  - ✦ TextOnly: train transformer without visual information
  - ✦ Conventional MMT models (e.g., DCCN (Lin et al. 2020), GMNMT (Yin et al. 2020), and Gated Fusion (Wu et al. 2021)
  - ✦ Text only inputs during test time ImagiT (Long et al. 2021), UVR-NMT (Fang et al. 2022)
- Evaluation Metrics:
  - ✦ BLEU - n-gram matching with ground truth
  - ✦ METOR - considers semantic similarity between generated text and ground truth

Results

Multi30k

Method	Model	EN → DE				EN → FR			
		Test2016	Test2017	MSCOCO	Average	Test2016	Test2017	MSCOCO	Average
Transformer-Base	T	32.0 ± 0.9	23.3 ± 0.8	21.3 ± 0.9	25.5 ± 0.9	59.7 ± 0.2	52.1 ± 0.1	42.4 ± 0.6	51.4 ± 0.3
Transformer-Small	T	38.2 ± 0.4	28.8 ± 0.4	25.8 ± 0.3	30.9 ± 0.4	58.4 ± 0.4	50.9 ± 0.3	41.6 ± 0.4	50.3 ± 0.4
	V	39.4 ± 0.3	31.7 ± 0.2	**27.9 ± 0.3**	33.0 ± 0.3	**60.5 ± 0.1**	52.3 ± 0.7	43.1 ± 0.3	52.0 ± 0.4
	VM	**39.6 ± 0.3**	**31.8 ± 0.2**	**27.9 ± 0.3**	**33.1 ± 0.3**	**60.5 ± 0.2**	**52.4 ± 0.6**	**43.4 ± 0.2**	**52.1 ± 0.3**
Transformer-Tiny	T	39.7 ± 0.3	31.7 ± 0.5	28.4 ± 0.2	33.3 ± 0.3	60.9 ± 0.5	53.7 ± 0.4	44.4 ± 0.2	53.0 ± 0.4
	V	**41.9 ± 0.2**	**34.0 ± 0.3**	30.3 ± 0.3	**35.4 ± 0.3**	62.3 ± 0.2	**55.1 ± 0.3**	**45.7 ± 0.2**	**54.4 ± 0.2**
	VM	**41.9 ± 0.2**	**34.0 ± 0.3**	**30.4 ± 0.4**	**35.4 ± 0.3**	**62.4 ± 0.3**	55.0 ± 0.3	**45.7 ± 0.4**	**54.4 ± 0.3**

Table 1: **BLEU score on Multi30K**. T: Baseline text-only transformer; V: **VALHALLA** model with hallucinated visual representations; VM: **VALHALLA** model with ground-truth visual representations. Please refer to supplementary material for METEOR score comparisons.

Results

Multi30k cont:

Method	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multimodal Machine Translation												
Gumbel-Attention [31]	39.2	57.8	31.4	51.2	26.9	46.0	–	–	–	–	–	–
CAP-ALL [29]	39.6	57.5	33.0	52.2	27.6	46.4	60.1	74.3	52.8	68.6	44.3	62.6
GMNMT [69]	39.8	57.6	32.2	51.9	28.7	47.6	60.9	74.9	53.9	69.3	–	–
DCCN [30]	39.7	56.8	31.0	49.9	26.7	45.7	61.2	76.4	54.3	70.3	45.4	65.0
VALHALLA (M)	**41.9**	**68.7**	**34.0**	**62.5**	**30.4**	**57.2**	**62.4**	**81.4**	**55.0**	**76.4**	**45.7**	**71.0**
Gated Fusion [66]	42.0	67.8	33.6	61.9	29.0	56.1	61.7	81.0	54.8	76.3	44.9	70.5
VALHALLA (M)	**42.6**	**69.3**	**35.1**	**62.8**	**30.7**	**57.6**	**63.1**	**81.8**	**56.0**	**77.1**	**46.4**	**71.3**
Text-Only Machine Translation												
$VMMT_F$ [7]	37.7	56.0	30.1	49.9	25.5	44.8	–	–	–	–	–	–
UVR-NMT [74]	36.9	–	28.6	–	–	–	58.3	–	48.7	–	–	–
ImagiT [32]	38.5	55.7	32.1	52.4	28.7	48.8	59.7	74.0	52.4	68.3	45.3	65.0
VALHALLA	**41.9**	**68.8**	**34.0**	**62.5**	**30.3**	**57.0**	**62.3**	**81.4**	**55.1**	**76.4**	**45.7**	**70.9**
RMMT [66]	41.4	68.0	32.9	61.7	30.0	56.3	62.1	81.3	54.4	76.1	44.5	70.2
VALHALLA	**42.7**	**69.3**	**35.1**	**62.8**	**30.7**	**57.5**	**63.1**	**81.8**	**56.0**	**77.1**	**46.5**	**71.4**

Table 2: **Comparison with state-of-the-art multimodal and text-only translation methods on Multi30K**. **VALHALLA** hallucinates visual representations from text-only inputs, while **VALHALLA (M)** uses ground-truth visual tokens at test time. Results in gray are computed with model averaging over 10 latest checkpoints. **VALHALLA** establishes new state-of-the-art for machine translation on Multi30K.

Results

Multi30k cont:

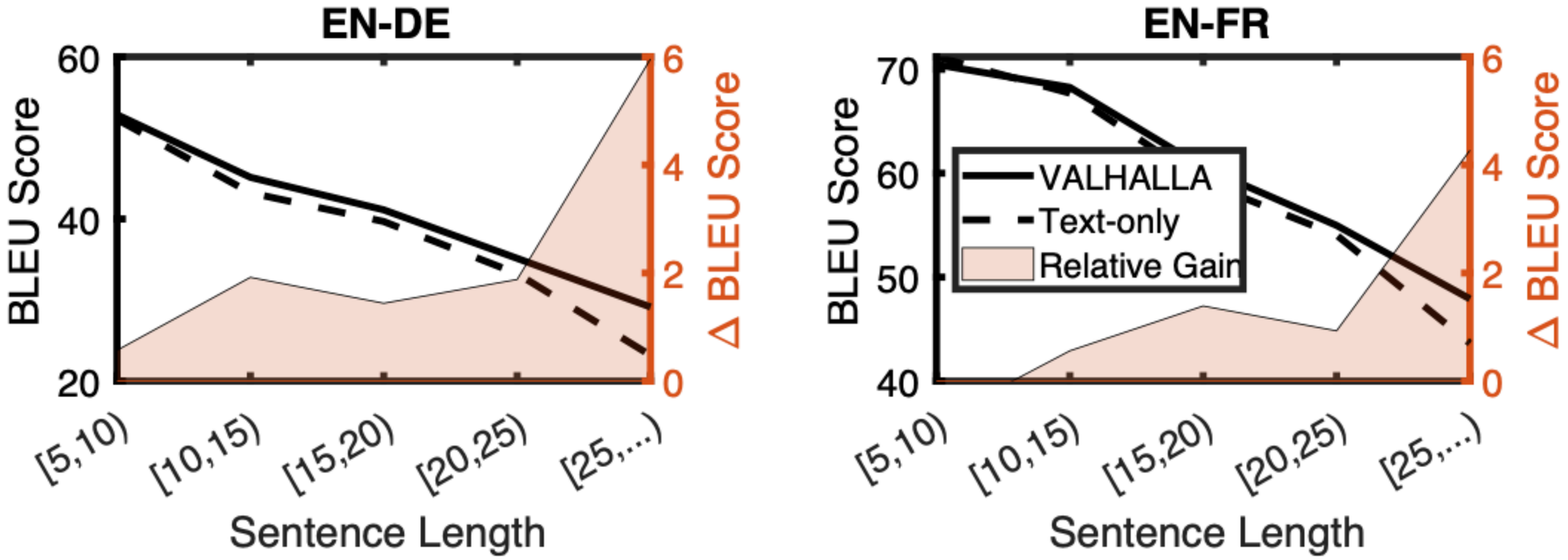


Figure 3: BLEU scores on different groups divided according to source sentence lengths on Multi30K Test2016 split.

Results

WIT/WMT

Method	Well-Resourced			Non-English		Under-Resourced		Average
	EN → DE	EN → ES	EN → FR	DE → ES	ES → FR	EN → RO	EN → AF	
Text-Only	16.0 ± 0.5	24.8 ± 0.8	16.1 ± 1.2	10.7 ± 0.2	16.2 ± 0.3	11.5 ± 0.7	10.8 ± 0.6	15.1 ± 0.6
UVR-NMT [74]	16.9 ± 0.2	26.4 ± 0.4	17.7 ± 0.3	10.9 ± 0.9	16.4 ± 0.6	12.5 ± 0.5	11.6 ± 1.7	16.1 ± 0.7
RMMT [66]	16.4 ± 0.3	24.8 ± 0.4	17.2 ± 1.6	11.0 ± 0.3	15.9 ± 0.7	9.9 ± 1.4	9.8 ± 1.0	15.0 ± 0.7
VALHALLA	**17.5 ± 0.4**	**27.5 ± 0.2**	**18.8 ± 0.2**	**11.3 ± 0.2**	**16.6 ± 0.8**	**14.4 ± 1.0**	**14.0 ± 0.5**	**17.2 ± 0.4**
VALHALLA (M)	17.4 ± 0.4	**27.5 ± 0.2**	**18.8 ± 0.2**	**11.3 ± 0.2**	**16.6 ± 0.8**	**14.4 ± 1.0**	**14.0 ± 0.4**	**17.2 ± 0.4**

Table 3: **BLEU score on WIT**. Please refer to supplementary material for METEOR score comparisons.

Method	Visual Data	Well-Resourced				Under-Resourced			
		EN → DE		EN → FR		EN → DE		EN → FR	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	–	27.1 ± 0.2	55.0 ± 0.1	39.1 ± 0.2	64.4 ± 0.1	16.7 ± 0.2	43.6 ± 0.2	25.9 ± 0.1	52.3 ± 0.3
UVR-NMT [74]	Multi30K	27.2 ± 0.2 (28.1)	55.3 ± 0.1	39.7 ± 0.2 (39.6)	64.9 ± 0.1	17.1 ± 0.1	44.1 ± 0.3	26.1 ± 0.3	52.8 ± 0.3
RMMT [66]	Multi30K	24.5 ± 0.2	52.8 ± 0.1	35.3 ± 0.0	61.2 ± 0.1	15.7 ± 0.2	41.9 ± 0.4	24.2 ± 0.3	50.7 ± 0.3
VALHALLA	Multi30K	**28.0 ± 0.1**	56.0 ± 0.1	**40.0 ± 0.1**	**65.2 ± 0.1**	17.6 ± 0.1	**44.8 ± 0.1**	**26.9 ± 0.2**	53.2 ± 0.2
	WIT	**28.0 ± 0.1**	**56.1 ± 0.1**	39.9 ± 0.1	65.1 ± 0.1	**17.7 ± 0.2**	44.7 ± 0.1	26.8 ± 0.0	**53.3 ± 0.1**
VALHALLA (M)	Multi30K	**28.0 ± 0.0**	56.0 ± 0.1	39.9 ± 0.1	65.0 ± 0.1	**17.7 ± 0.1**	**44.8 ± 0.2**	**26.9 ± 0.2**	**53.3 ± 0.3**
	WIT	27.9 ± 0.1	56.0 ± 0.2	39.8 ± 0.2	65.0 ± 0.1	**17.7 ± 0.2**	**44.8 ± 0.1**	26.8 ± 0.1	**53.3 ± 0.1**

Table 4: **Results on WMT2014**. UVR-NMT results in brackets are reported by the original paper.

Results

Translation Under Limited Textual Context

Progressive Masking: replaces all except the first K words of source sentence

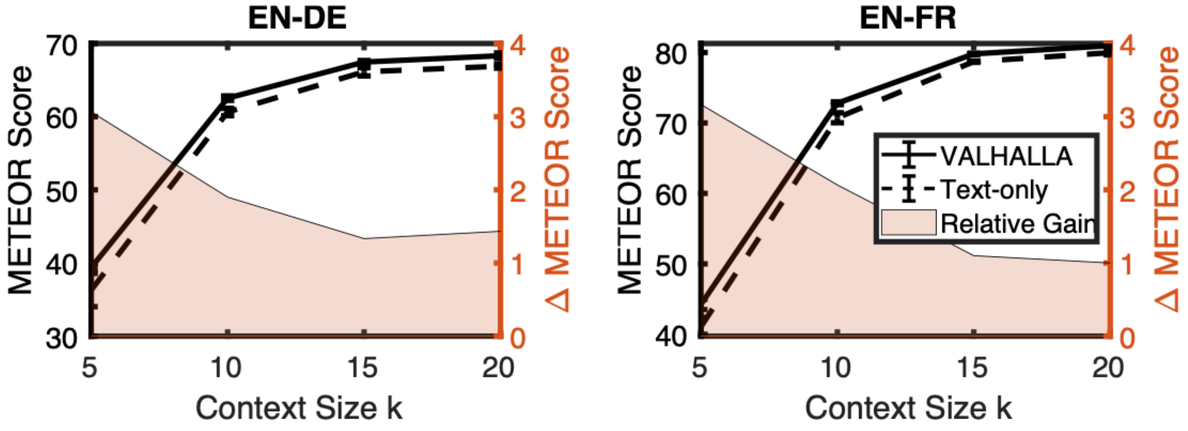


Figure 4: **Evaluation with Progressive Masking.** All results use METEOR scores on Multi30K Test2016 split.

Visual Entity Masking: randomly replaces visually grounded phrases

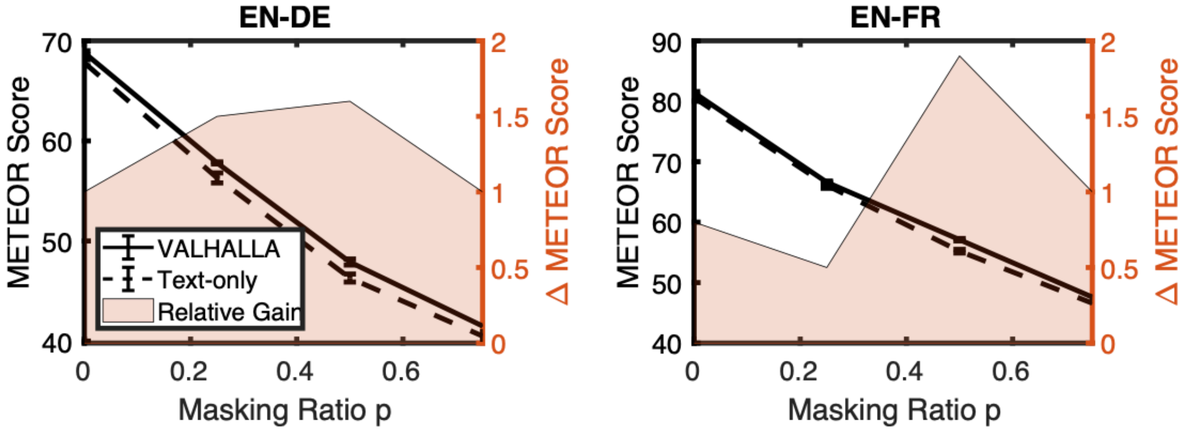


Figure 5: **Evaluation with Entity Masking.** All results use METEOR scores on Multi30K Test2016 split.

Ablation Analysis

- **Discrete vs Continuous Representation**
- **Visual Encoder Parameters**
- **Image Retrieval**

Backbone	Discrete Embedding	External Pretraining	Aggregation	EN-DE	EN-FR
CLIP RN-50	✗	CLIP	Gating	38.0	58.8
ResNet-50	✗	ImageNet	Gating	38.8	59.1
			Concatenation	38.3	60.0
VQGAN VAE	✓	None	Concatenation	**39.6**	**60.5**

(a) Discrete and continuous visual encoder backbones, evaluated with Transformer-Small on Multi30K Test2016 split.

Encoder Layers	Visual Token Length	EN-DE	EN-FR
4	$16^2 = 256$	13.5 ± 7.2	54.3 ± 0.4
5	$8^2 = 64$	36.3 ± 0.2	60.3 ± 0.2
6	$4^2 = 16$	**39.6 ± 0.3**	**60.5 ± 0.1**

(b) Visual encoder depths, evaluated with Transformer-Small on Multi30K Test2016.

Visual Data	Image Retrieval	EN-DE	EN-FR
Multi30K	✗	16.5 ± 0.3	26.2 ± 0.1
	✓	**17.6 ± 0.1**	**26.9 ± 0.2**
WIT	✗	16.6 ± 0.2	26.1 ± 0.3
	✓	**17.7 ± 0.2**	**26.8 ± 0.0**

(c) Training on WMT under-resourced tasks *without* image retrieval.

Ablation Analysis

- **Randomized Visual Tokens**
  - Decrease performance when used random visual tokens
- **Loss Hyperparameters**

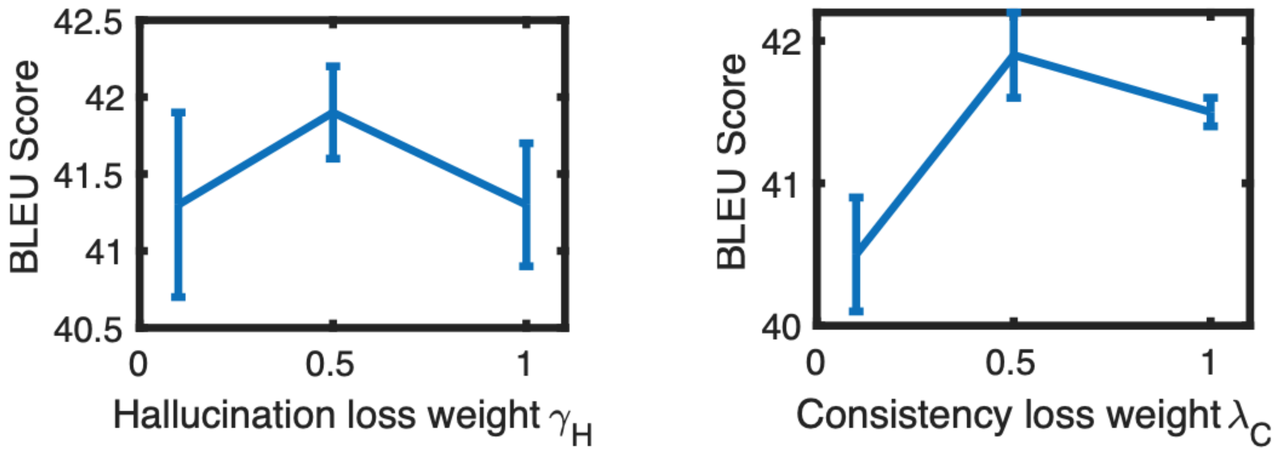


Figure 7: Influence of loss weights γ_H and λ_C of (13) on translation performance, measured on Multi30K EN→DE task.

Strengths

- Outperforms other baselines in different settings with extensive experiments
  - In longer sentence the performance gain is much better than short sentence
- No prior work is done on large scale multilingual WIT dataset
  - Proposed a new benchmark in seven language pairs under three settings.
- Generate semantically more informed translation compared to text only models in progressive masking scenario
- A detailed study indicates the visual information improves the performance and translation quality

Weaknesses

- The improvement in non-English language translation is marginal compared to text only baseline in WIT dataset
  - English-centric bias in image text pair
- The concept of hallucination is not novel
- Requires high quality image for the hallucination networks
- Training Complexity:
  - Careful tuning of hyperparameters
  - Sensitive towards consistency loss hyperparameter

Future Works

- Can extend the work with other modalities - video, audio
- More focus on the under represented/ non English language translation tasks

Discussions

- Robustness of the model training since hyperparameter sensitive
- Training complexity
    - Pre-extract VQGAN VAE tokens