MODULATING EARLY VISUAL PROCESSING BY LANGUAGE

HARM DE VRIES*[1], FLORIAN STRUB*[2], JEREMIE MARY[2], HUGO LAROCHELLE [1,3], OLIVIER PIETQUIN[1,5], AARON COURVILLE[1,4]

[1]MILA, UNIVERSITÉ OF MONTRÉAL - [2]UNIV. LILLE, CNRS, CENTRALE LILLE, INRIA, UMR 9189 CRIStAL - [3]GOOGLE BRAIN - [4]CIFAR - [5]DEEPMIND

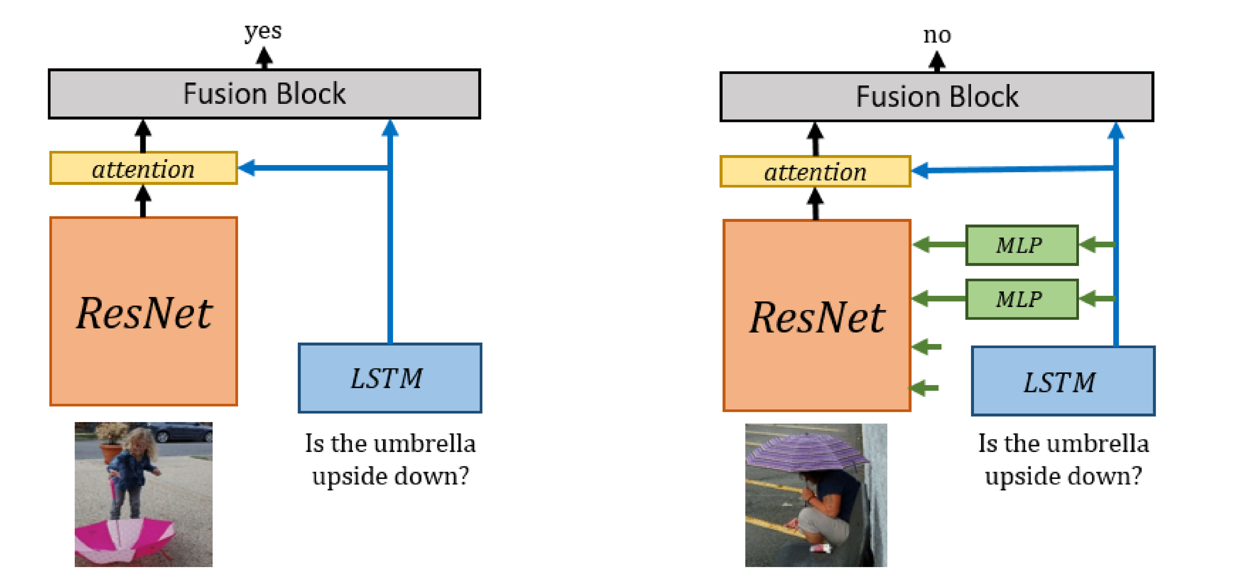
RETHINKING LANGUAGE-VISION TASKS

Premise In language-vision tasks (VQA, image captioning, instruction following), the classic pipeline processes the visual and linguistic inputs independently before fusing them into a single representation. This joint-embedding is then used to solve the task at hand.

Claim Linguistic input should modulate the visual processing from the very beginning to more effectively fuse both modalities and to obtain a better joint-embedding.

Solution We introduce Conditional Batch Normalization as a modulation mechanism to alter activations of a pre-trained ResNet conditioned on a language embedding.

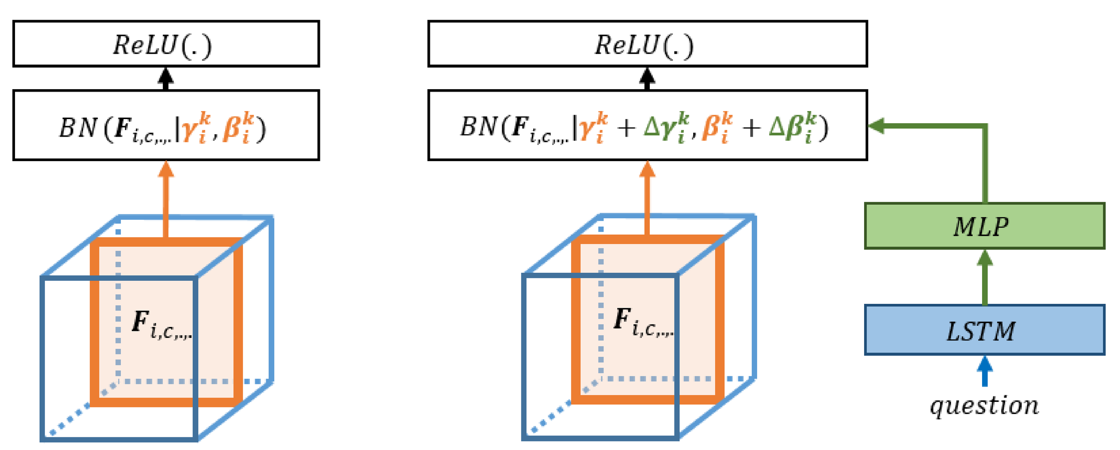
Results We show strong improvements on the VQA and GuessWhat?! datasets and find that early visual modulation is beneficial.



(Left) Classic language-vision tasks pipeline. (Right) Our proposed approach.

CONDITIONAL BATCHNORM

The idea is to condition the affine scaling parameters of a Batch Normalization (BN) layer, γ and β, with an external input $\boldsymbol{e}_q$. When applied to a pre-trained convnet, we predict a change $\Delta\beta_c$ and $\Delta\gamma_c$ from pre-initialized BN parameters.



We refer to $F_{i,c,w,h}$ as feature map of the i^{th} input sample of the c^{th} feature map at location (w,h). Given a mini-batch $\mathcal{B} = \{F_{i,\cdot,\cdot,\cdot}\}_{i=1}^{N}$ of N examples, **Conditional Batch Normalization** (CBN) normalizes the feature maps at training time as follows:

$$\boldsymbol{\Delta\beta} = MLP(\boldsymbol{e}_q) \qquad \boldsymbol{\Delta\gamma} = MLP(\boldsymbol{e}_q)$$

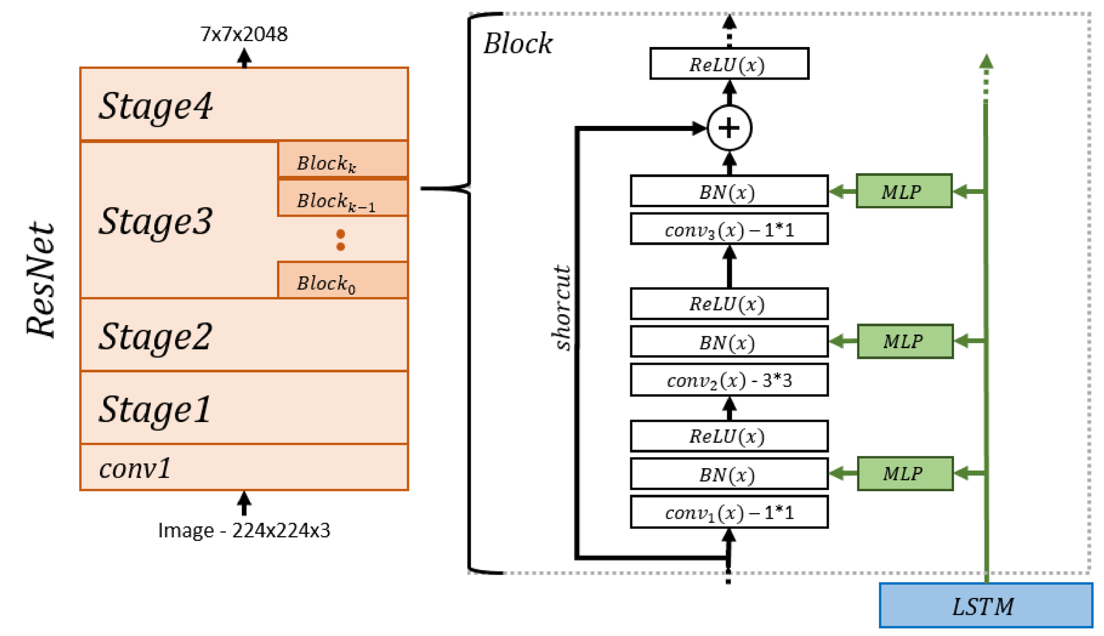
$$CBN(F_{i,c,h,w}) = (\gamma_c + \Delta\gamma_c)\frac{F_{i,c,w,h} - \mathrm{E}_{\mathcal{B}}[F_{\cdot,c,\cdot,\cdot}]}{\sqrt{\mathrm{Var}_{\mathcal{B}}[F_{\cdot,c,\cdot,\cdot}] + \epsilon}} + (\beta_c + \Delta\gamma_c),$$

CBN is a powerful method to modulate neural activations as it enables an external embedding to manipulate entire feature maps:

- by scaling them up or down if $\gamma_c > 0$
- by shifting them $\beta_c \neq 0$
- by shutting them off if $\gamma_c = 0$

MODULATING RESNET

In order to modulate the visual pipeline, we condition the BN parameters of a pre-trained ResNet on a language embedding obtained from a recurrent network. We train end-to-end but we stress that we freeze all ResNet parameters, including γ and β, during training.



We apply CBN to a pretrained ResNet-50, leading to the MODulatEd Residual Network (MODERN). To verify that the gains from MODERN are not coming from increased model capacity, we include two baselines with more capacity:

- **Ft Stage 4**: when finetuning the layers of stage 4 of ResNet-50
- **Ft BN**: when finetuning all β and γ parameters of ResNet-50, while freezing all its weights.

VQA

What color are her eyes?	Brown
What is the mustache made of?	Banana
Is he a boy?	Yes

GUESSWHAT?!

Is it a vase?	Yes
Is it on the left corner?	No
Is it the turquoise and purple one?	Yes

VQA RESULTS

Although MODERN can be combined with any existing VQA architecture, in this work we plug it into a original VQA architecture with either a classic spatial attention mechanism or a 2-glimpse attention mechanism [2]. Models are trained on the training set with early stopping on the validation set and *accuracies* are reported on test-dev set.

	Answer type	Yes/No	Number	Other	**Overall**
224x224	Baseline	79.45%	36.63%	44.62%	58.05%
	Ft Stage 4	78.37%	34.27%	43.72%	56.91%
	Ft BN	80.18%	35.98%	46.07%	58.98%
	MODERN	81.17%	37.79%	48.66%	60.82%
448x448	MRN [2] with ResNet-50	80.20%	37.73%	49.53%	60.84%
	MRN [2] with ResNet-152	80.95%	38.39%	50.59%	61.73%
	MCB [1] with ResNet-50	60.46%	38.29%	48.68%	60.46%
	MCB [1] with ResNet-152	-	-	-	62.50%
	MODERN	81.38%	36.06%	51.64%	62.16%
	MODERN + MRN [2]	82.17%	38.06%	52.29%	**63.01**%

CBN applied to	Val. accuracy
$\emptyset$	56.12%
Stage 4	57.68%
Stages 3 − 4	58.29%
Stages 2 − 4	58.32%
All	**58.56**%

GUESSWHAT?! RESULTS

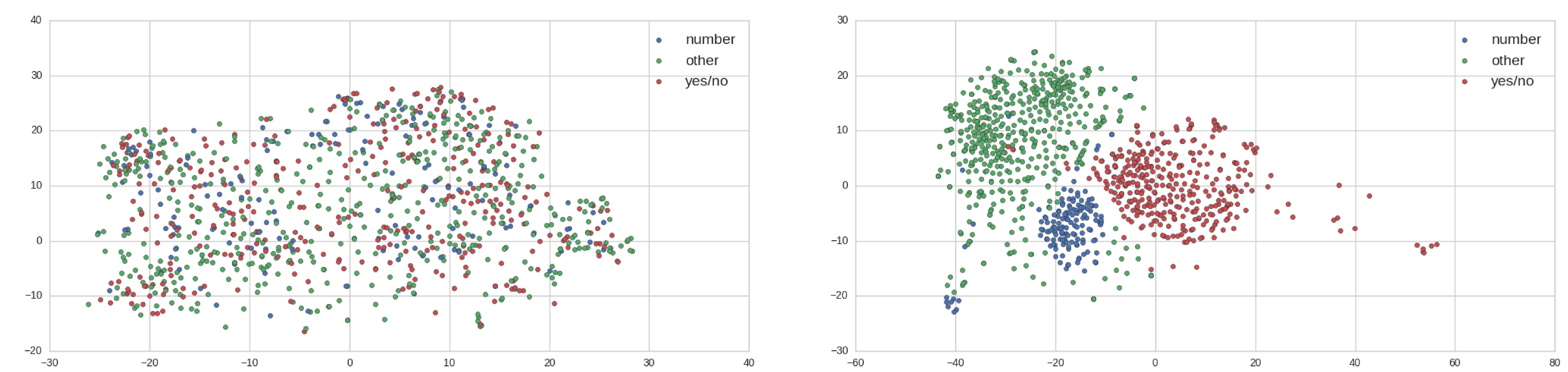
We use the oracle model as defined in the original GuessWhat?! paper with the (modulated) cropped object features, the object category, its spatial location and the question embedding as input. Models are trained on the training set with early stopping on the validation sets and *error* are reported on test set.

	Raw ResNet	ft stage4	Ft BN	CBN
Crop	29.92%	27.48%	27.94%	25.06%
Crop + Spatial + Cat.	22.55%	22.68%	22.42%	**19.52**%
Spatial + Category	21.5%			

CBN applied to	Test error
$\emptyset$	29.92%
Stage 4	26.42%
Stages 3 − 4	25.24%
Stages 2 − 4	25.31%
All	**25.06**%

MODULATED RESNET FEATURES DISENTANGLE ANSWER TYPES

For VQA, we show a t-SNE plot of 1000 raw (Left) and modulated (Right) ResNet-50 features.



REFERENCES

[1] Multimodal compact bilinear pooling for visual question answering and visual grounding. A. Fukui et Al. In *Proc. of EMNLP*, 2016.

[2] Hadamard product for low-rank bilinear pooling. J. Kim et Al. In *Proc. of ICLR*, 2017.