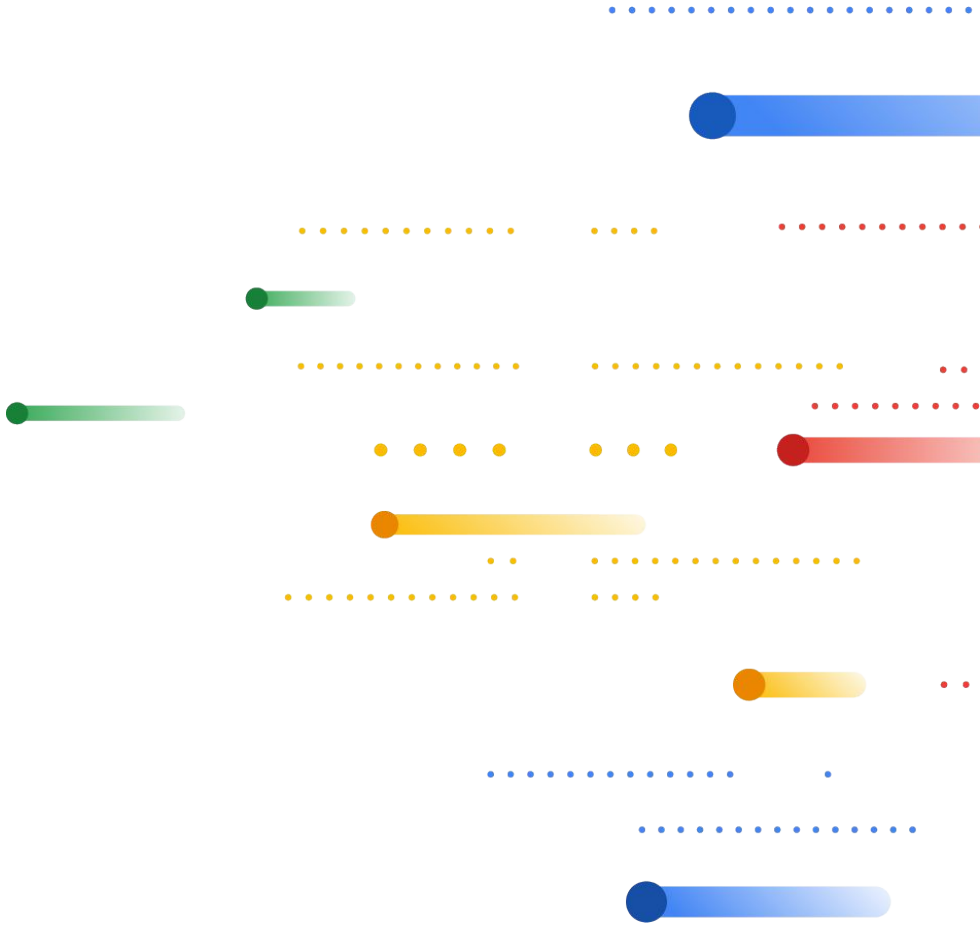
Google AI

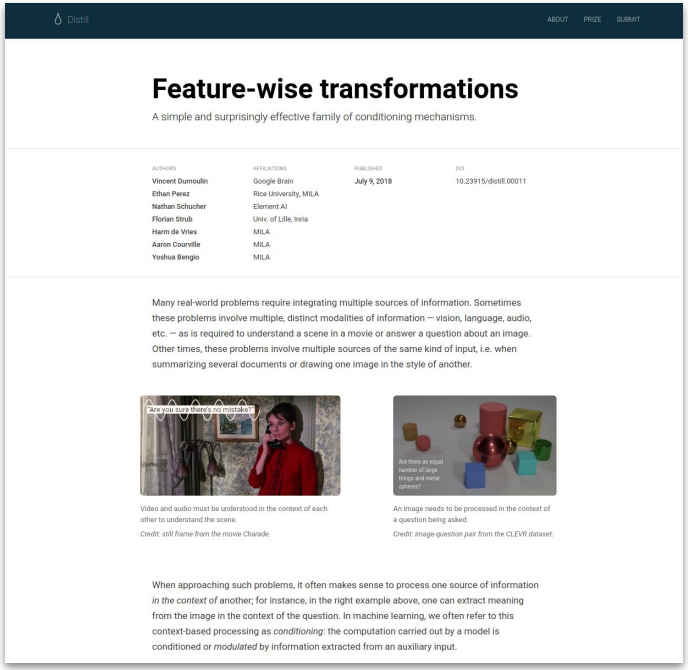
Feature-Wise Transformations

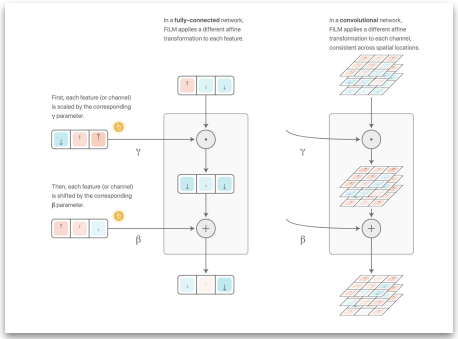
AMMI, May 31, 2019

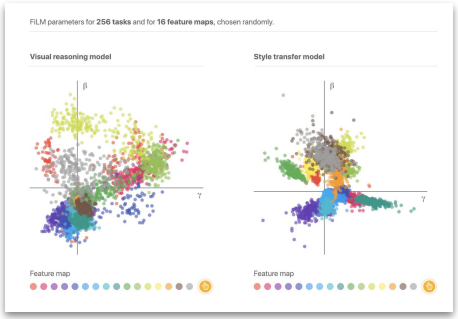


Vincent Dumoulin

Distill article

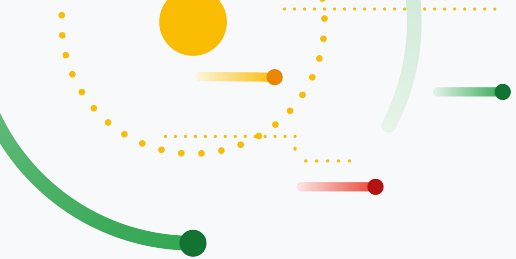







```
@article{dumoulin2018feature-wise,
  author = {Dumoulin, Vincent and
            Perez, Ethan and
            Schucher, Nathan and
            Strub, Florian and
            Vries, Harm de and
            Courville, Aaron and
            Bengio, Yoshua},
  title = {Feature-wise transformations},
  journal = {Distill},
  year = {2018},
  note = {https://distill.pub/2018/
          feature-wise-transformations},
  doi = {10.23915/distill.00011}
}
```


https://distill.pub/2018/feature-wise-transformations (Dumoulin et al., 2018)

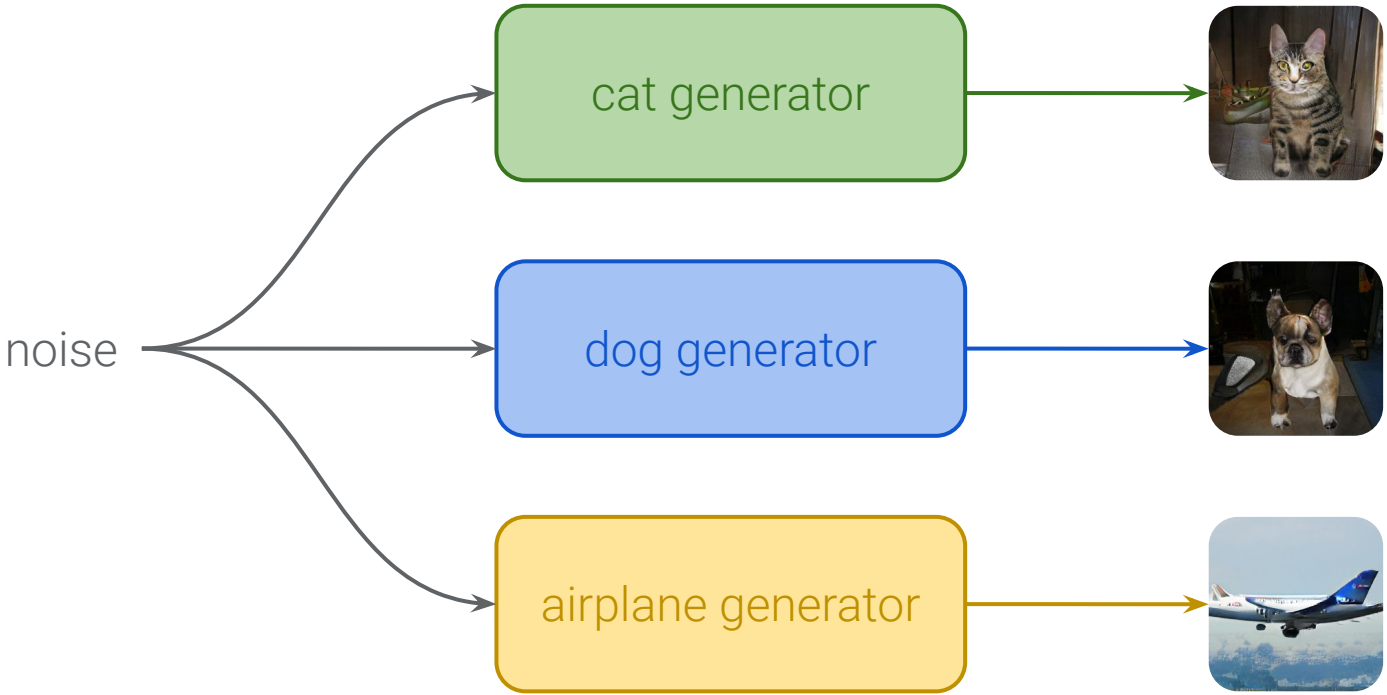


Google AI

Perspectives

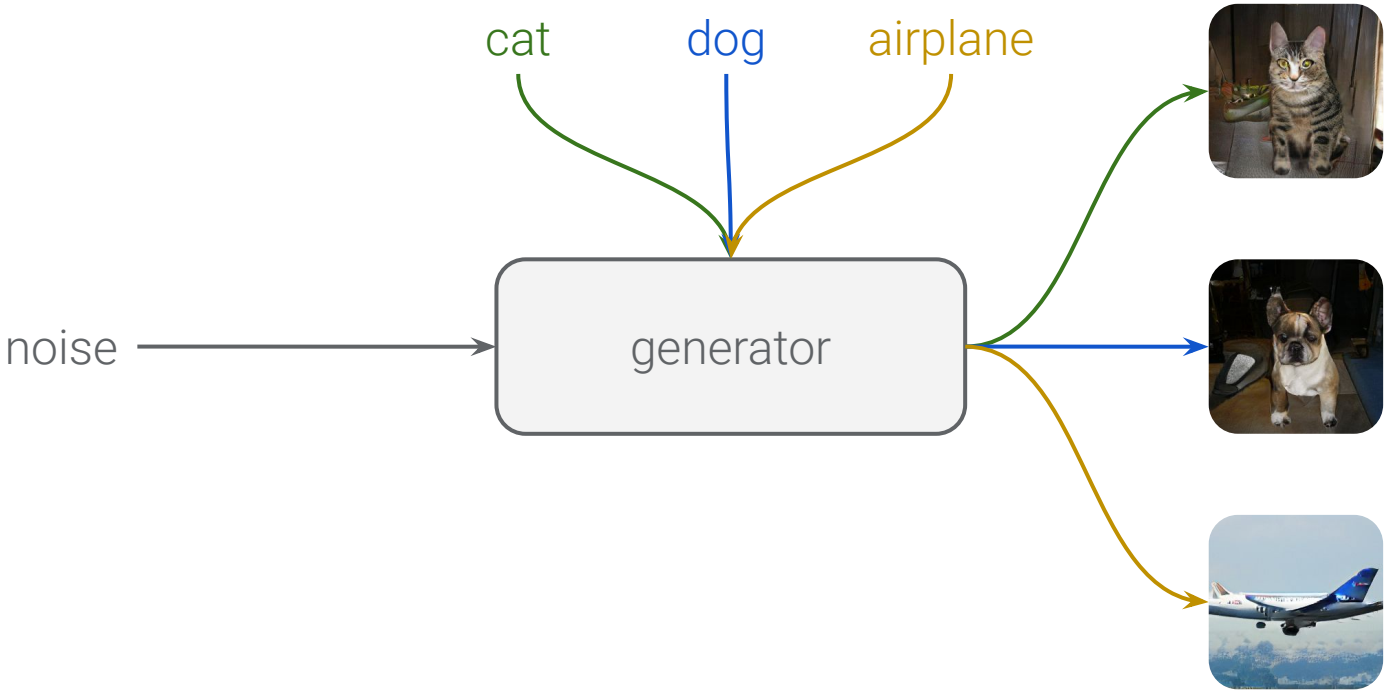
on learning problems

Class-conditional generative modeling



Samples drawn from a trained BigGAN model (Brock et al., 2019)

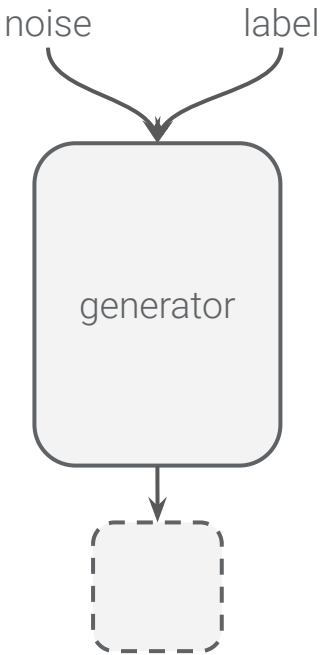
Class-conditional generative modeling



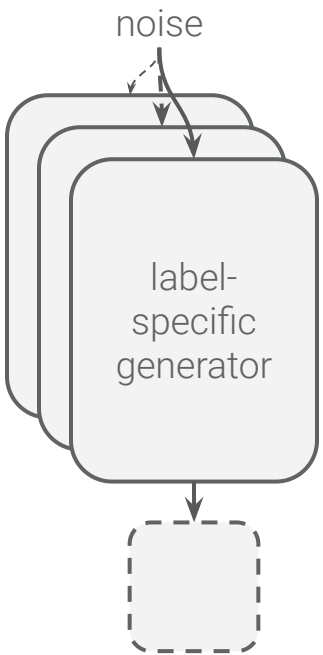
Samples drawn from a trained BigGAN model (Brock et al., 2019)

Learning perspectives

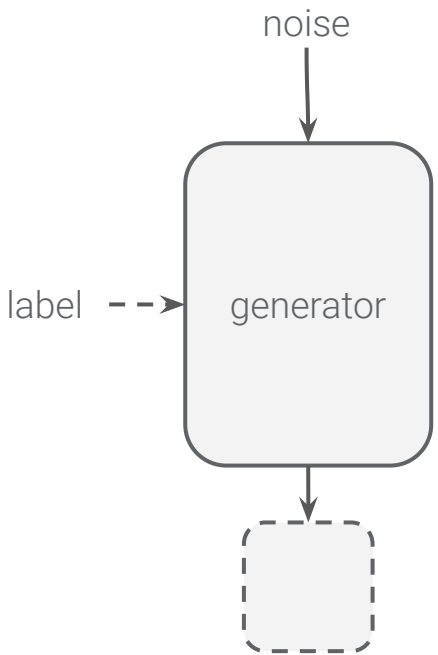
Modality fusion



Multi-task learning



Conditioning



Visual reasoning

Modality fusion perspective

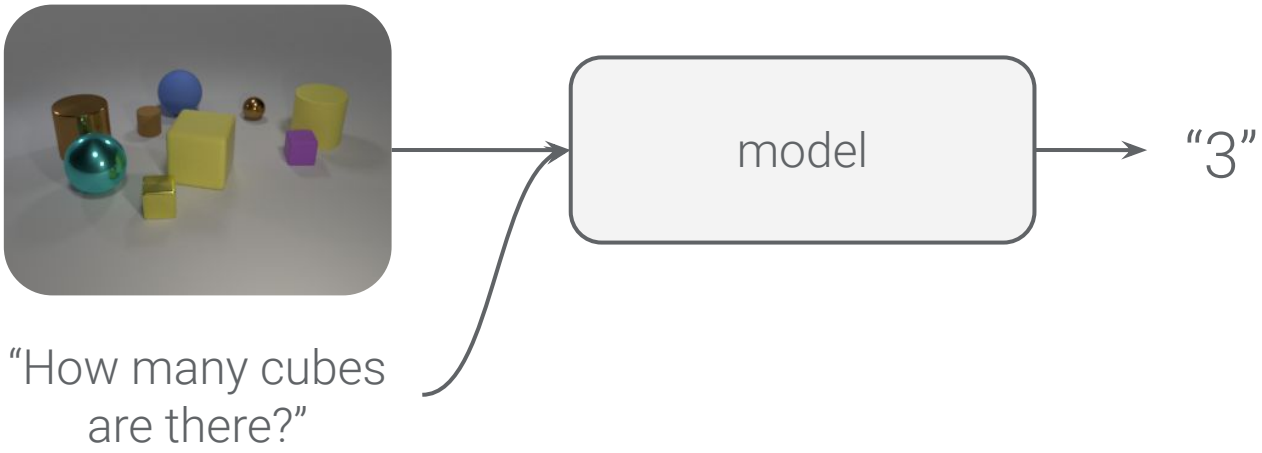


Image taken from the CLEVR dataset (Johnson et al., 2017)

Visual reasoning

Conditioning perspective

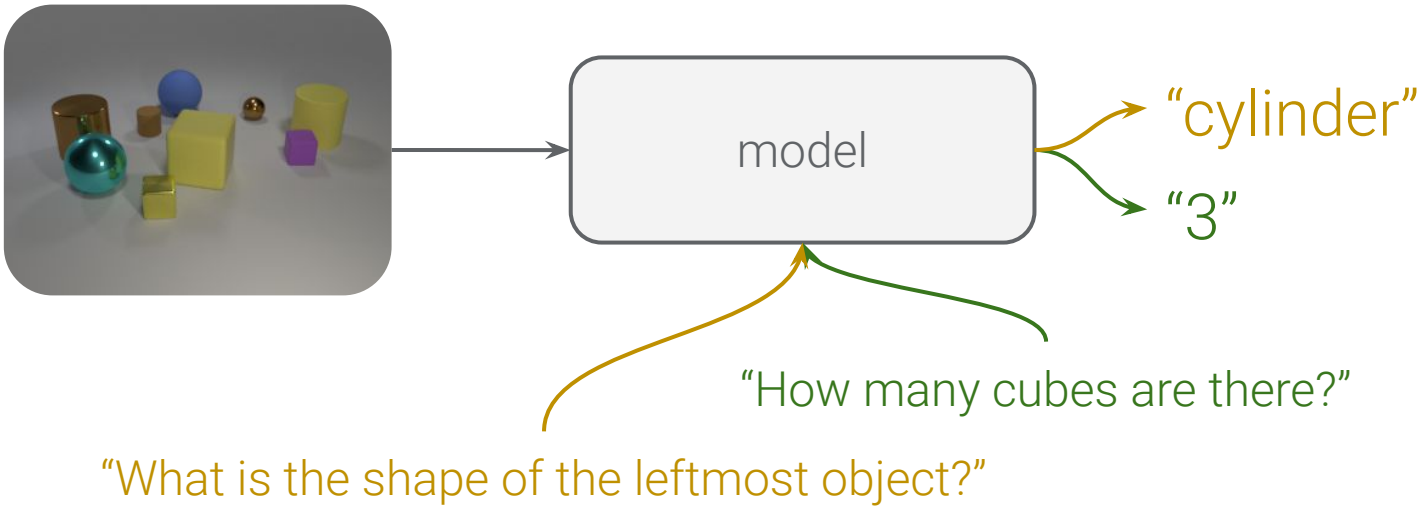
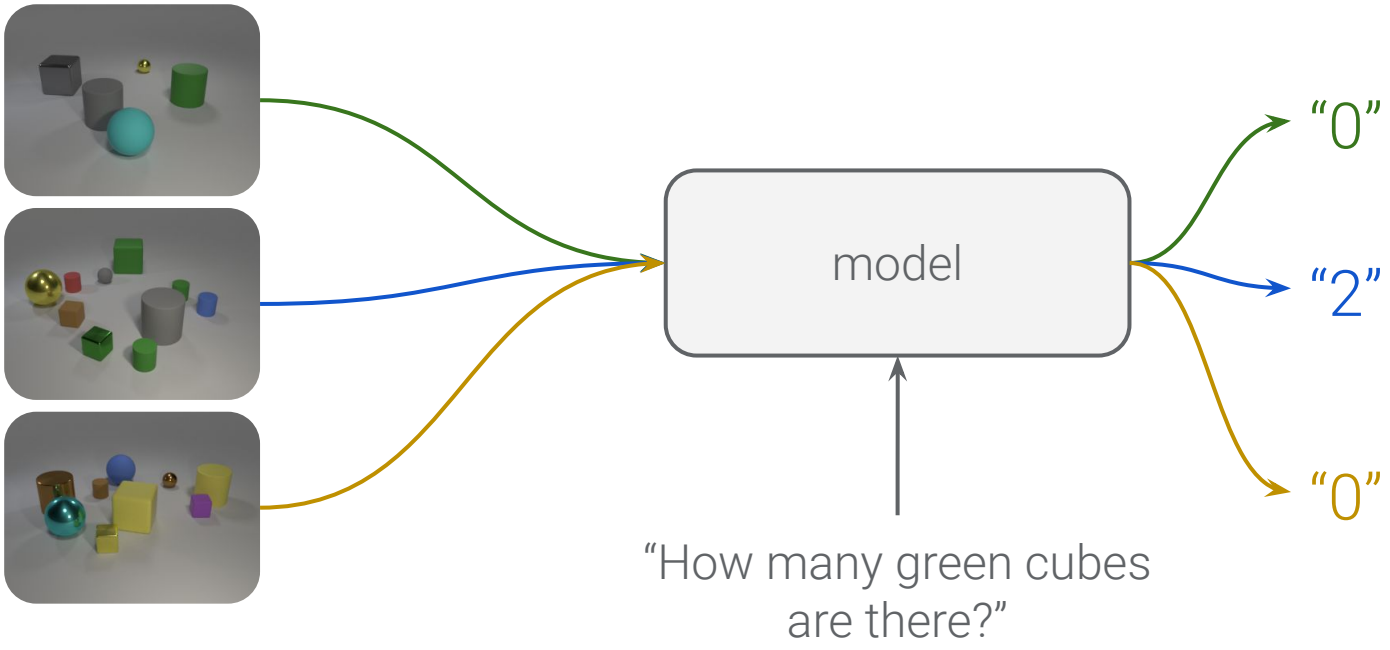


Image taken from the CLEVR dataset (Johnson et al., 2017)

Visual reasoning

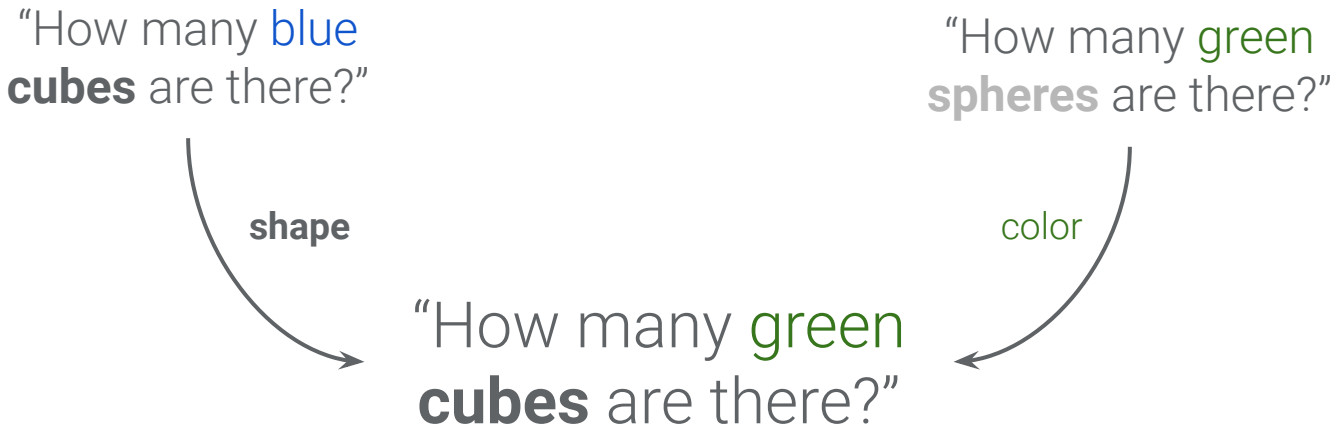
Multi-task learning perspective

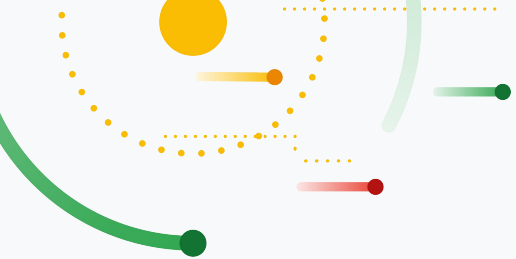


Images taken from the CLEVR dataset (Johnson et al., 2017)

Visual reasoning

Zero-shot learning perspective

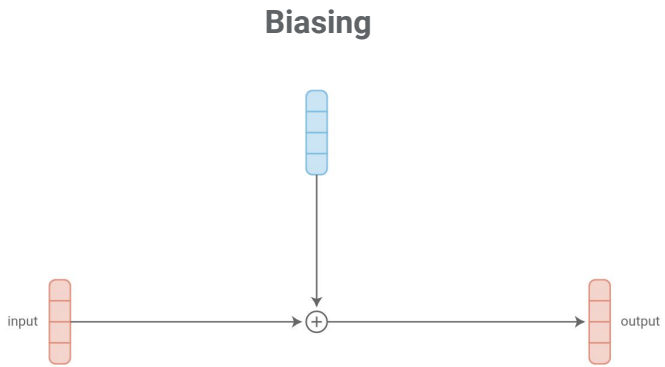


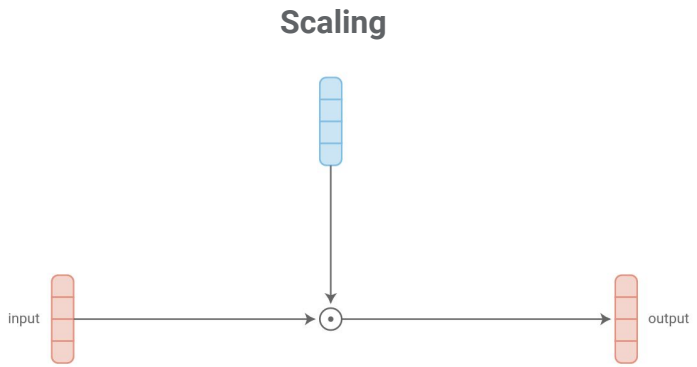


Feature-wise transformations

What are feature-wise transformations?

A transformation on an input feature vector — or stack of feature maps — which acts **independently** on individual features — or feature maps.

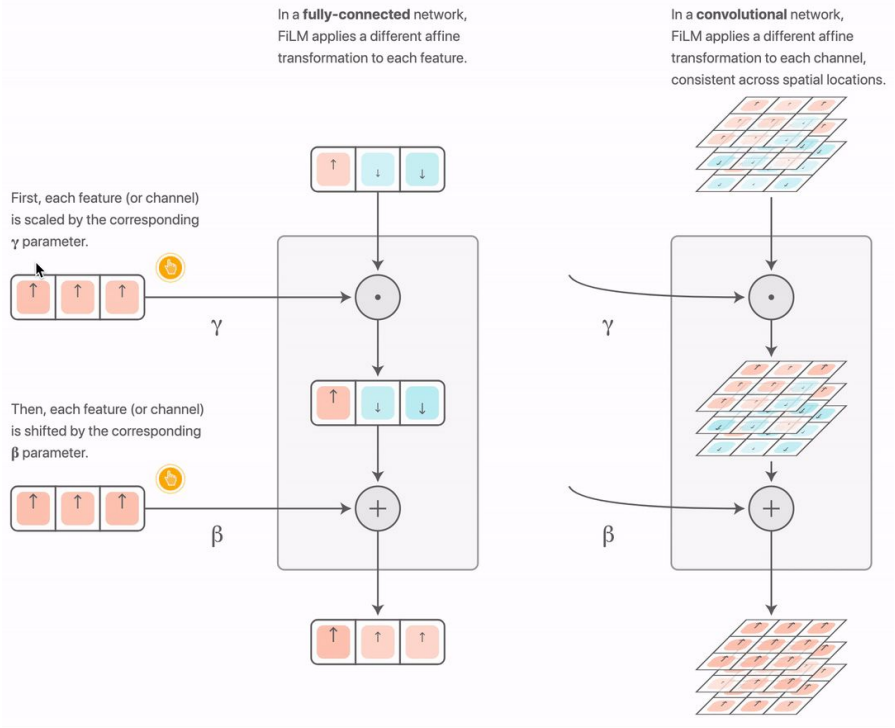




Figures from Dumoulin et al. (2018)

FiLM nomenclature

Feature-w**i**se **L**inear **M**odulation



$\gamma, \beta \in \mathbb{R}$ → **FiLM**

$\gamma = 1, \beta \in \mathbb{R}$ → **biasing**

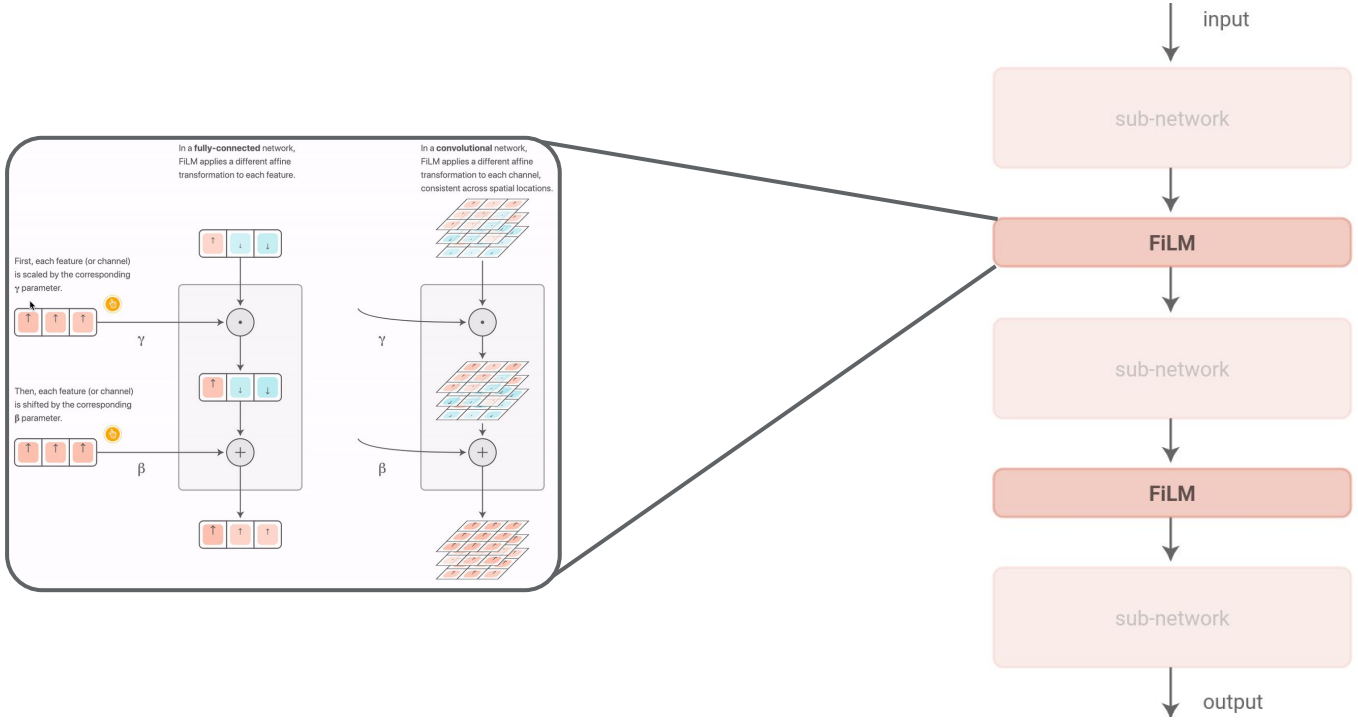
$\gamma \in \mathbb{R}, \beta = 0$ → **scaling**

$\gamma \in [0, 1], \beta = 0$ → **gating**

Figure from Dumoulin et al. (2018)

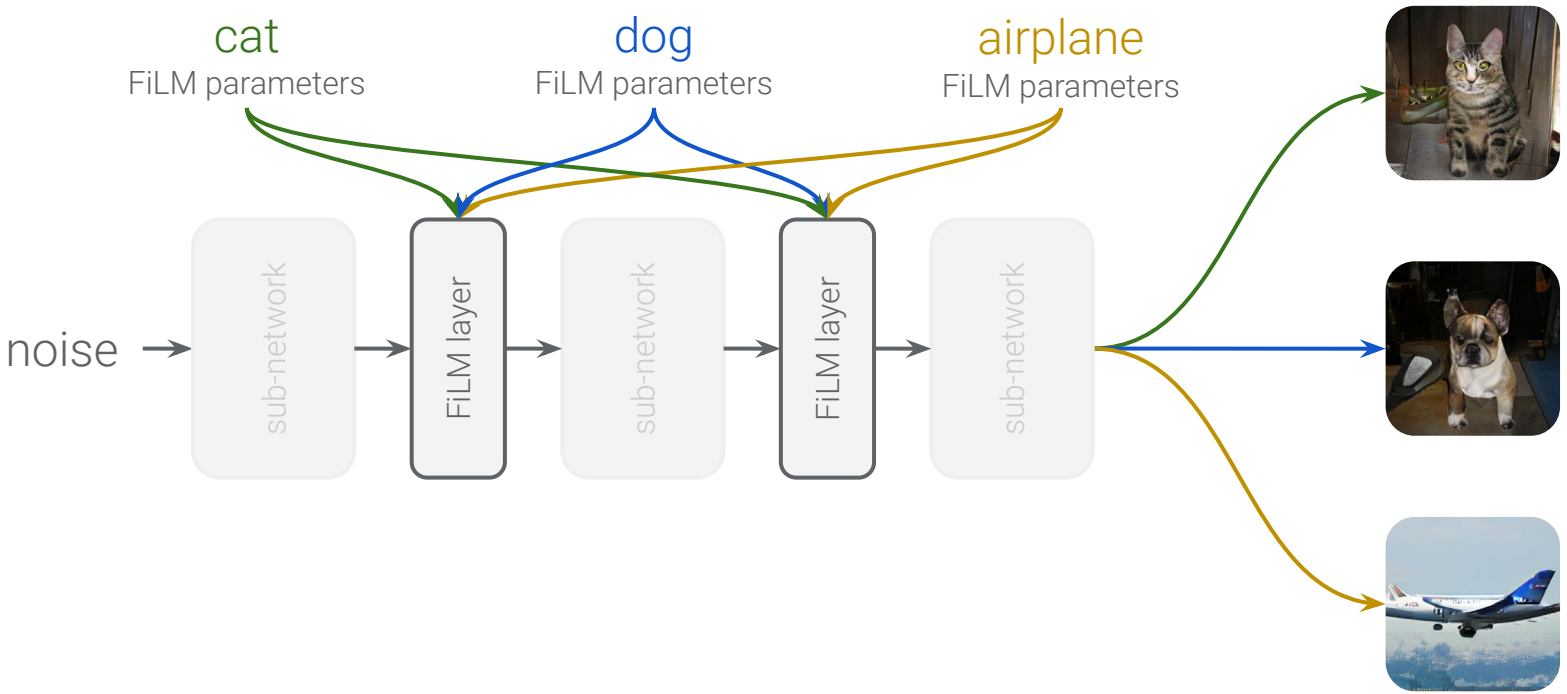
FiLM nomenclature

Feature-w**i**se **L**inear **M**odulation



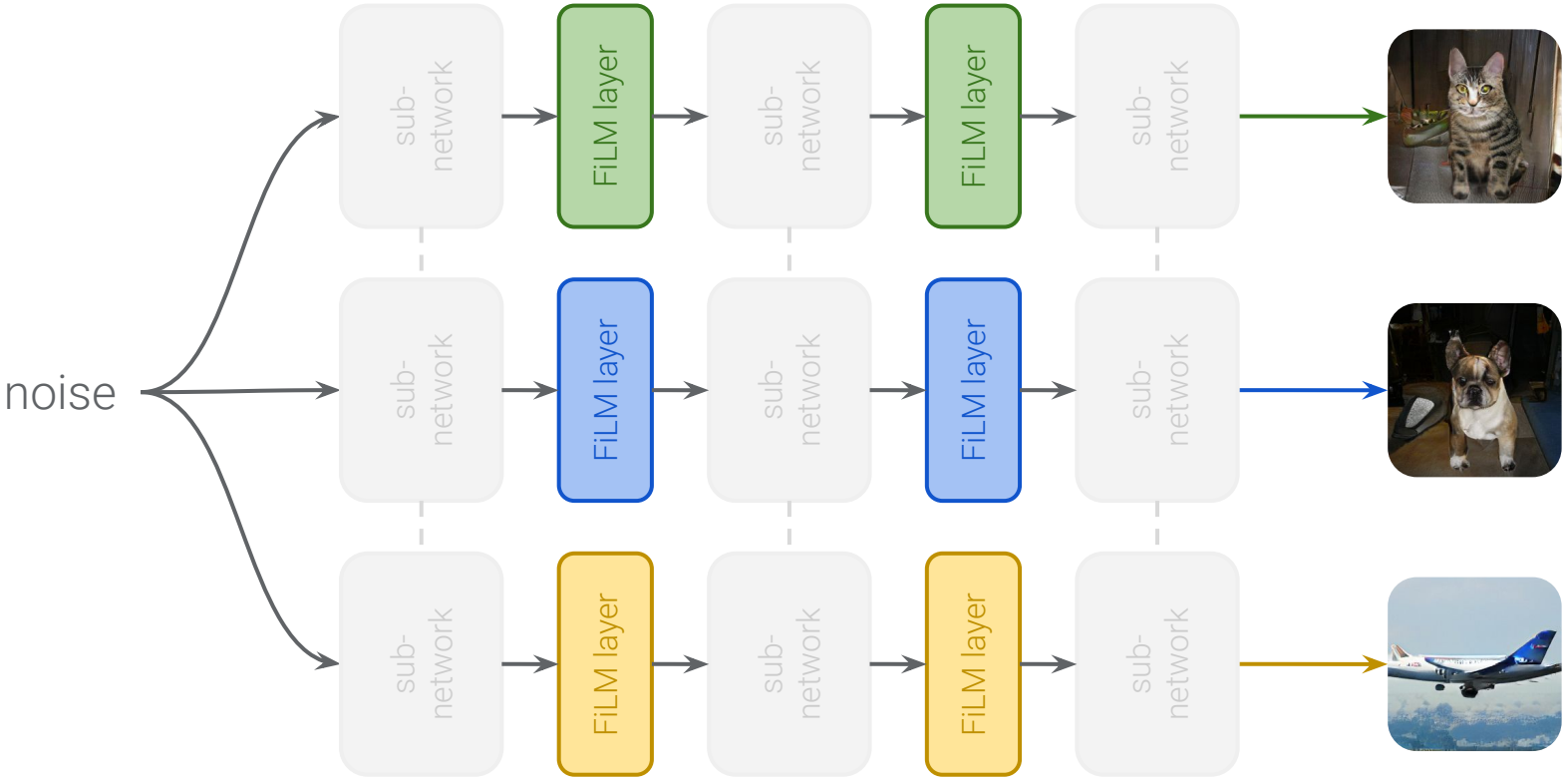
Figures from Dumoulin et al. (2018)

FiLM and class-conditional generative modeling



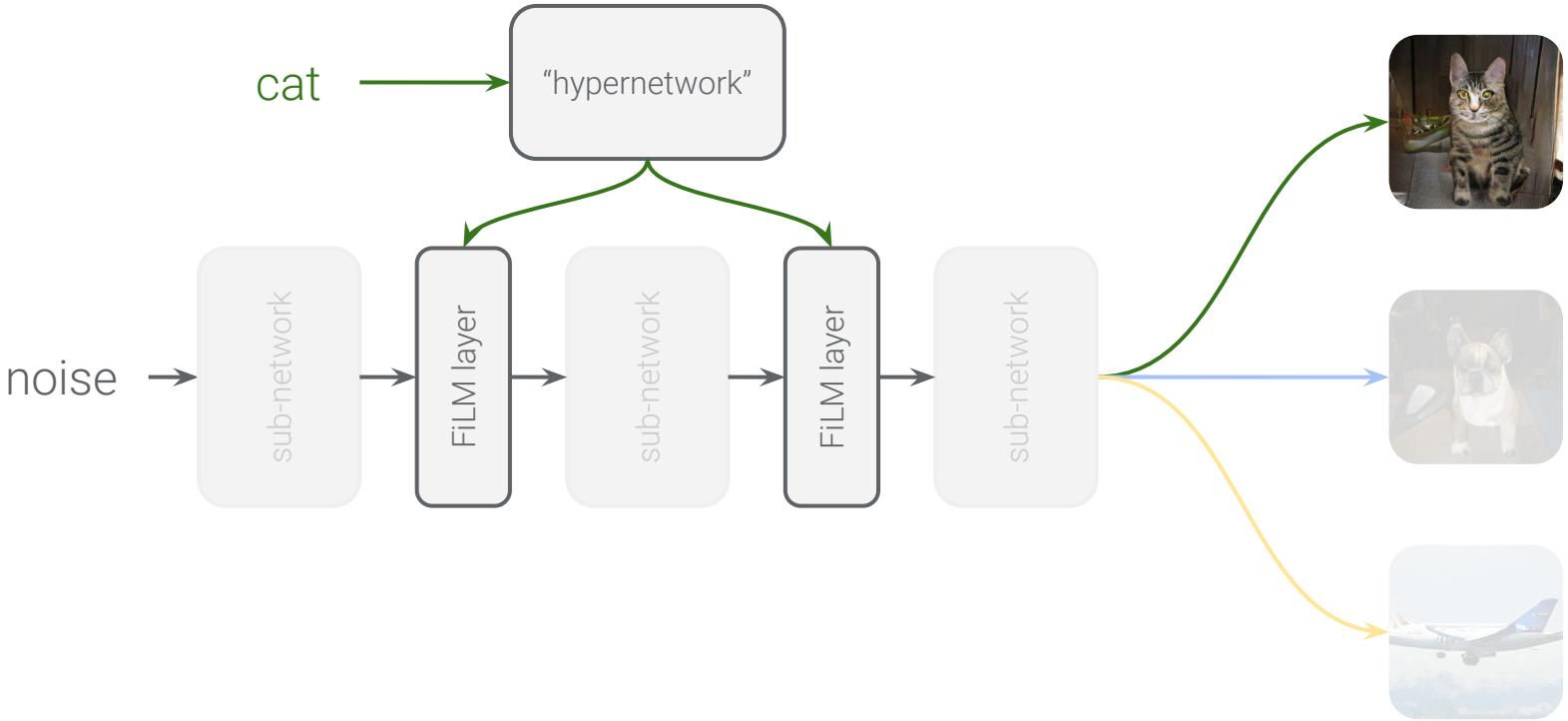
Samples drawn from a trained BigGAN model (Brock et al., 2019)

FiLM and class-conditional generative modeling



Samples drawn from a trained BigGAN model (Brock et al., 2019)

FiLM and class-conditional generative modeling



Samples drawn from a trained BigGAN model (Brock et al., 2019)

FiLM and visual reasoning

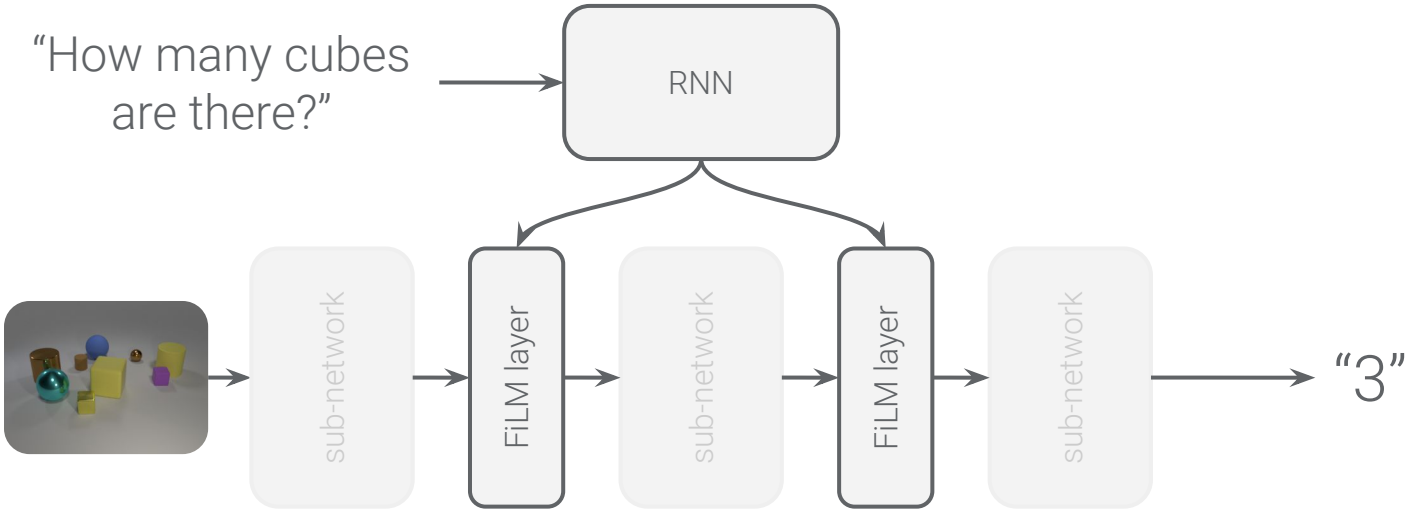


Image taken from the CLEVR dataset (Johnson et al., 2017)

General FiLM framework

The **FiLM generator** processes the conditioning information and produces parameters that describe how the target network should alter its computation.

Here, the **FiLM-ed network**'s computation is conditioned by two FiLM layers.

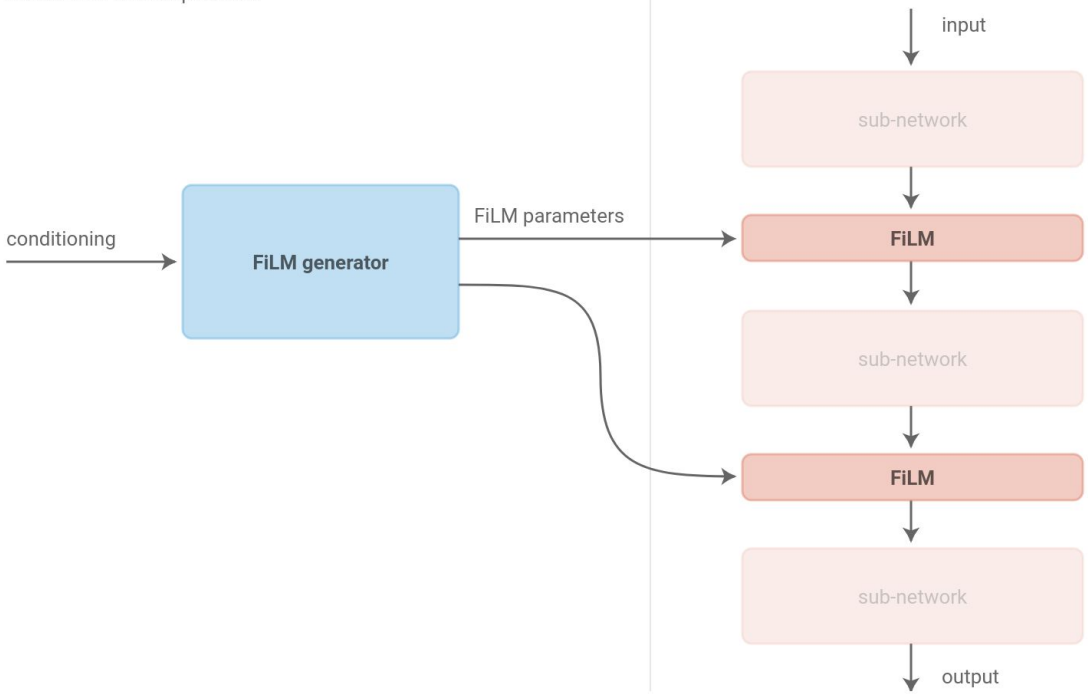
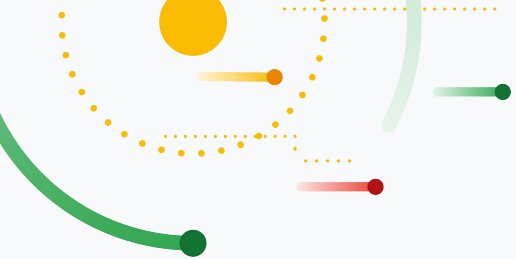


Figure from Dumoulin et al. (2018)



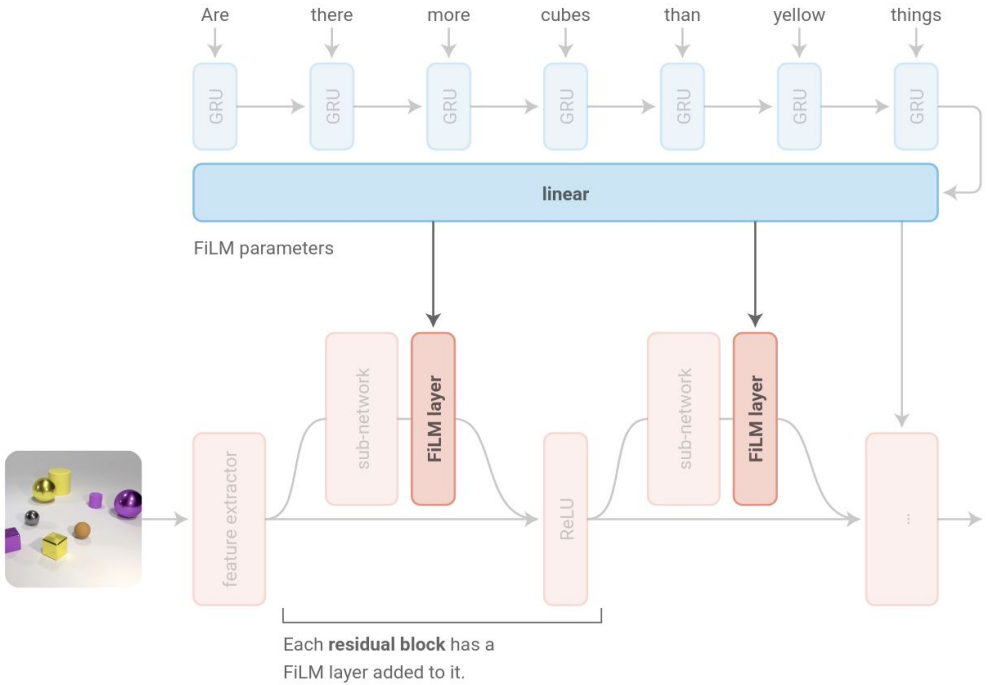
Google AI

Examples

FiLM: Visual Reasoning with a General Conditioning Layer

The **linguistic pipeline** acts as the FiLM generator.

FiLM layers in each residual block modulate the **visual pipeline**.



Perez et al. (2018); figure from Dumoulin et al. (2018)

Style transfer

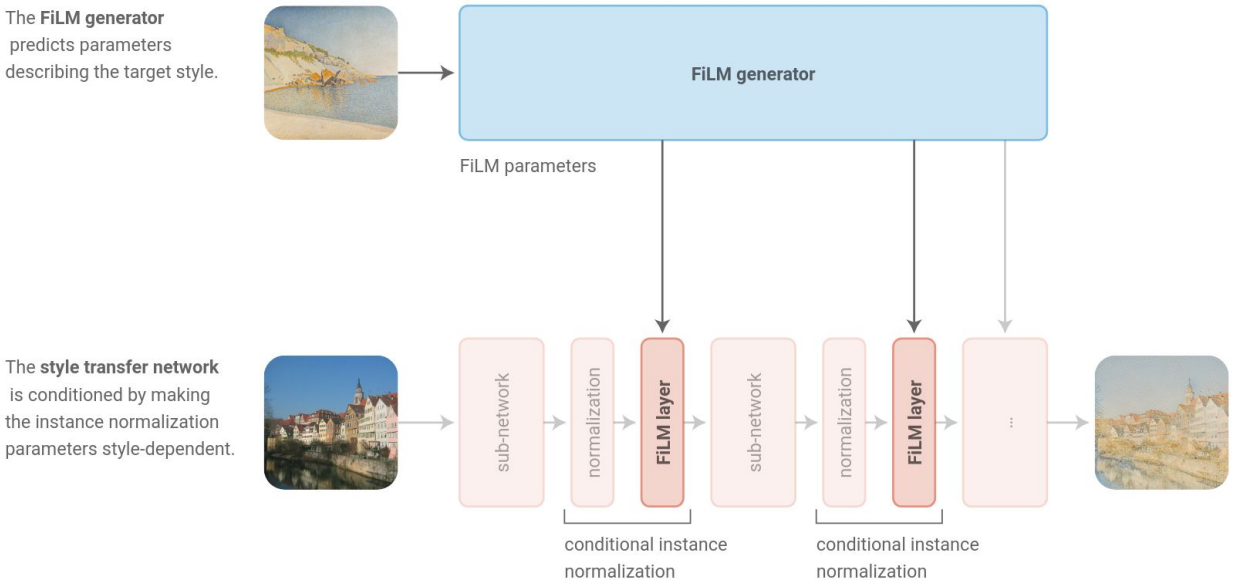


Figure from Dumoulin et al. (2018)

A learned representation for artistic style

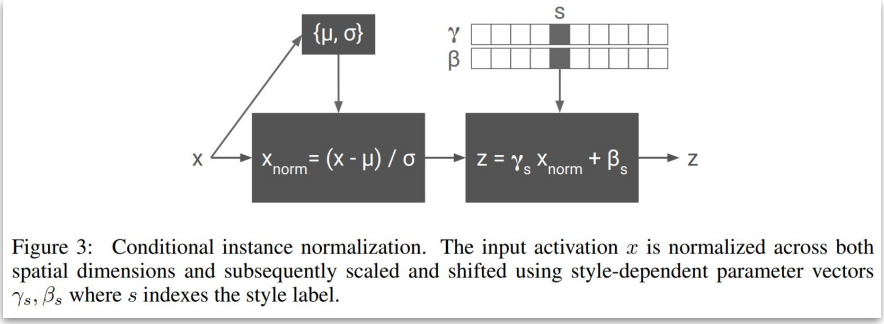
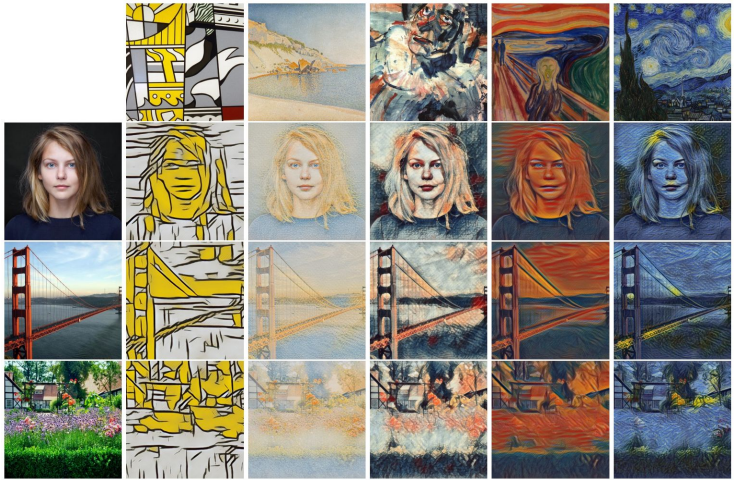
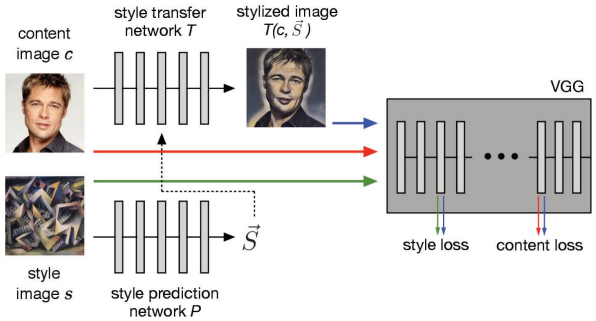


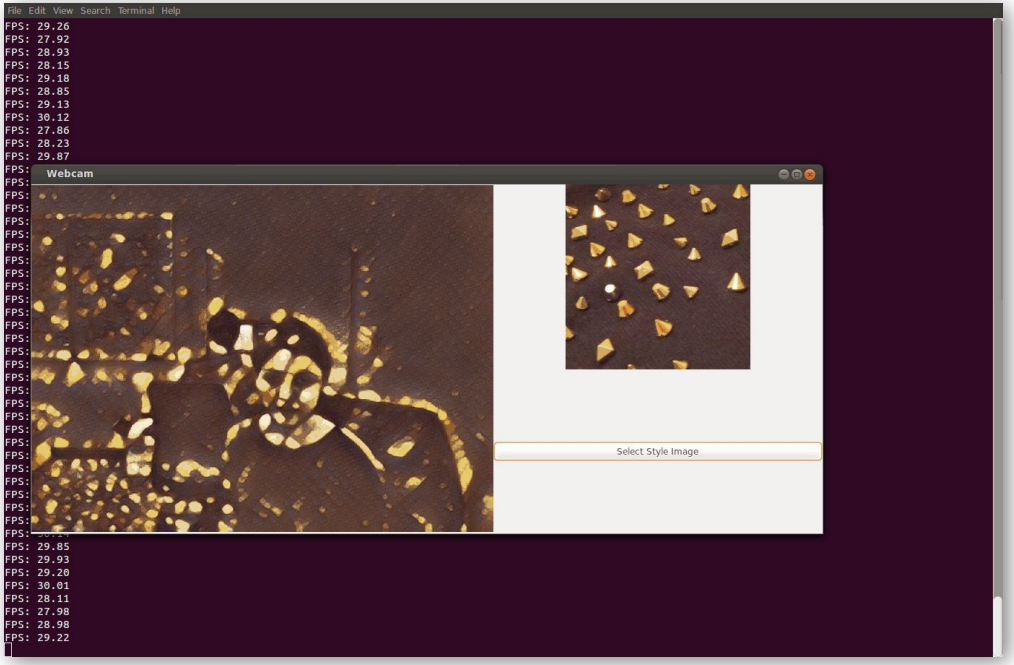
Figure 3: Conditional instance normalization. The input activation x is normalized across both spatial dimensions and subsequently scaled and shifted using style-dependent parameter vectors γ_s, β_s where s indexes the style label.



Dumoulin et al. (2017); figures from the article

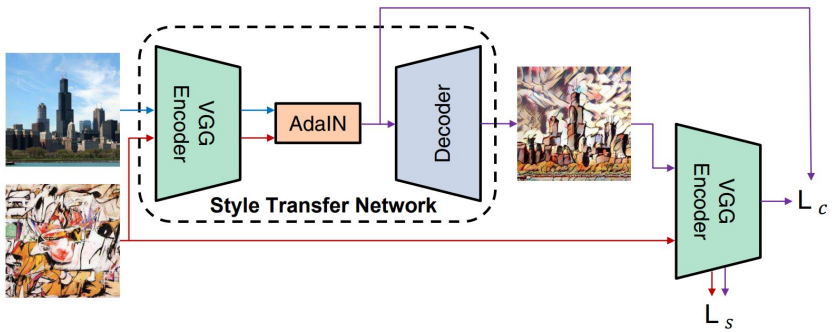
Exploring the structure of a real-time, arbitrary neural artistic stylization network

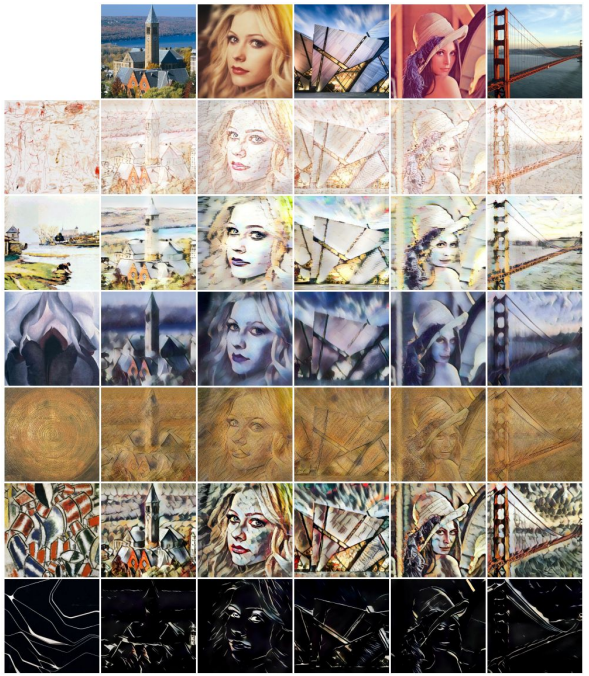




Ghiasi et al. (2017); figure from the article, video produced by the authors

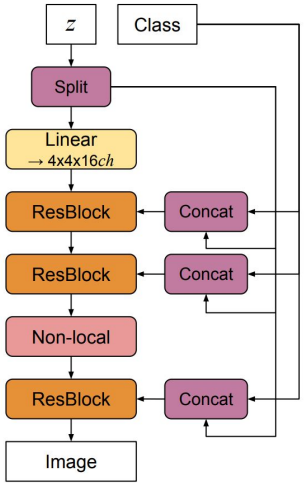
Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization (AdaIN)

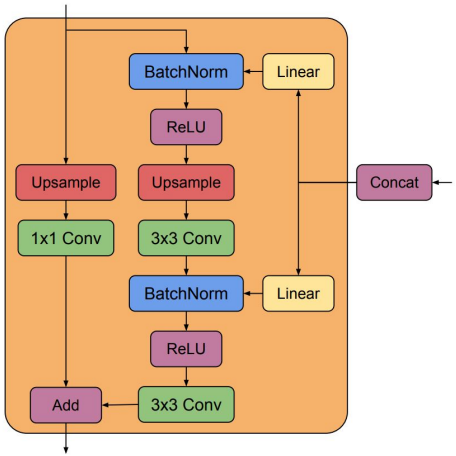




Huang & Belongie (2017); figures from the article

Large Scale GAN Training for High Fidelity Natural Image Synthesis (BigGAN)



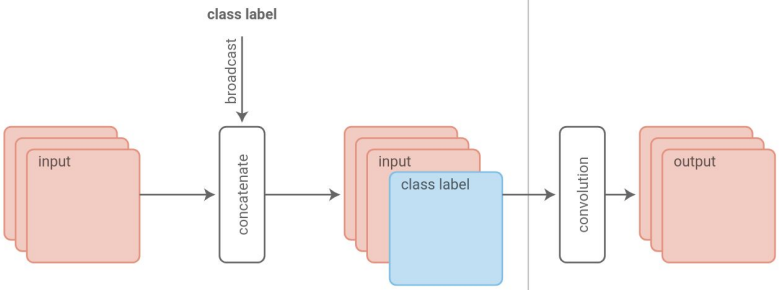


Brock et al. (2018); figures from the article

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (DCGAN)

Concatenation-based conditioning is used in the class-conditional DCGAN model. Each convolutional layer is concatenated with the broadcased label along the channel axis.

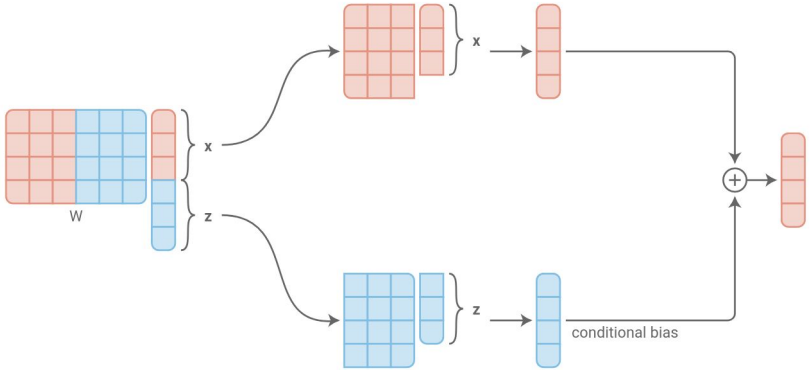
The resulting stack of feature maps is then **convolved** to produce the conditioned output.



Concatenation-based conditioning is equivalent to **conditional biasing**.

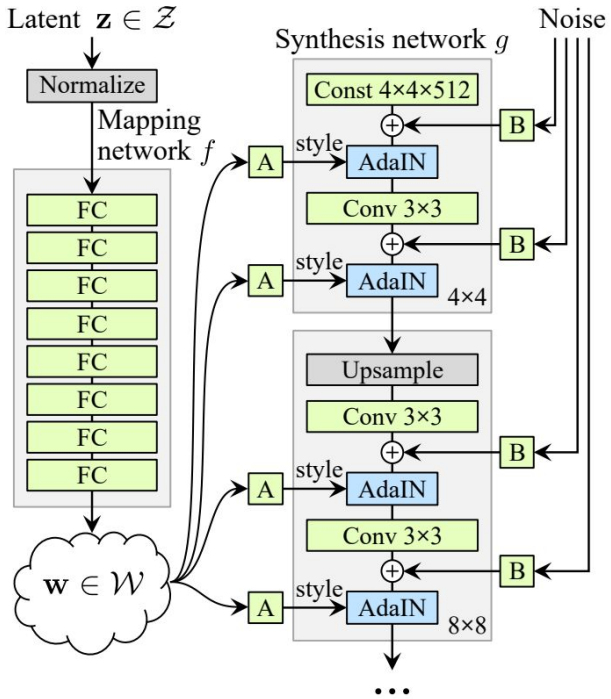
We can decompose the matrix-vector product into two matrix-vector subproducts.

We can then add the resulting two vectors. The **z**-dependent vector is a conditional bias.



Radford et al. (2016); figures from Dumoulin et al. (2018)

A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN)



Karras et al. (2019); figures from the article

Slimmable neural networks

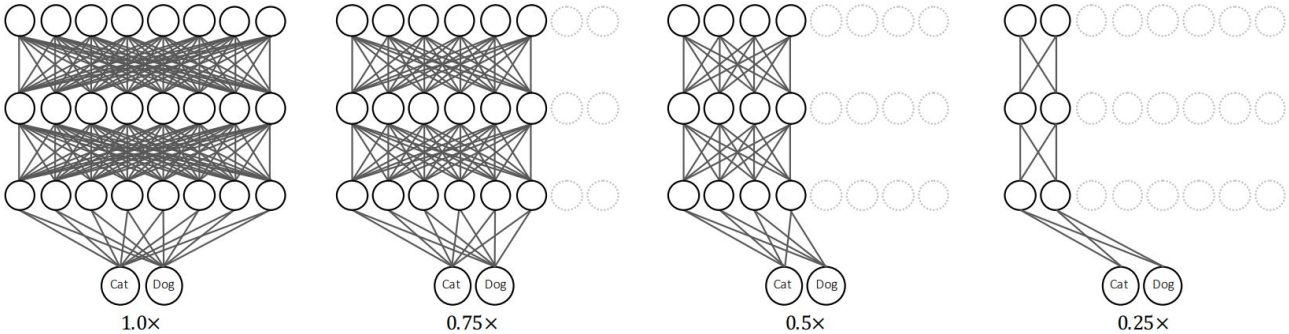
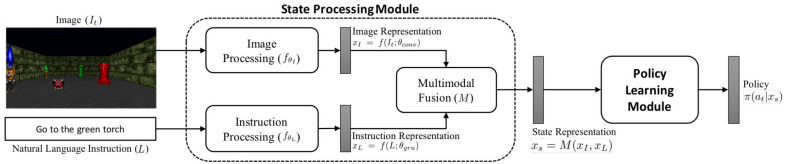
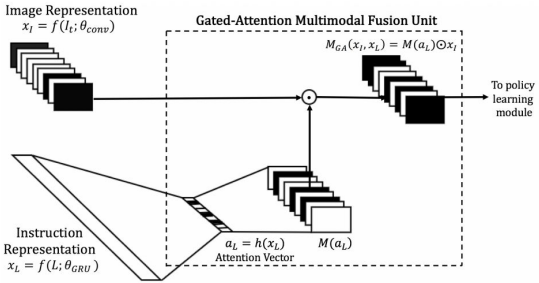


Figure 1: Illustration of slimmable neural networks. The same model can run at different widths (number of active channels), permitting instant and adaptive accuracy-efficiency trade-offs.

Yu et al. (2019); figure from the article

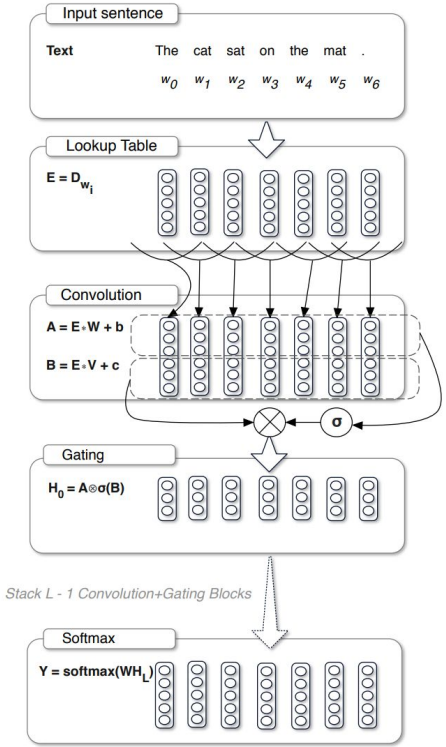
Gated-Attention Architectures for Task-Oriented Language Grounding





Chaplot et al. (2018); figures and video from the article

Language Modeling with Gated Convolutional Networks



Dauphin et al. (2018); figure from the article

Squeeze-and-Excitation Networks

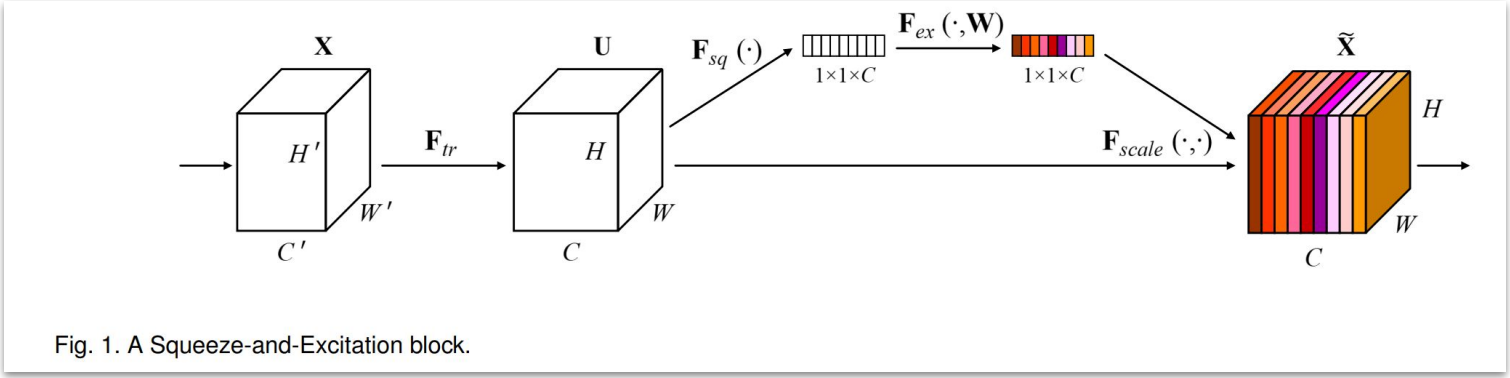
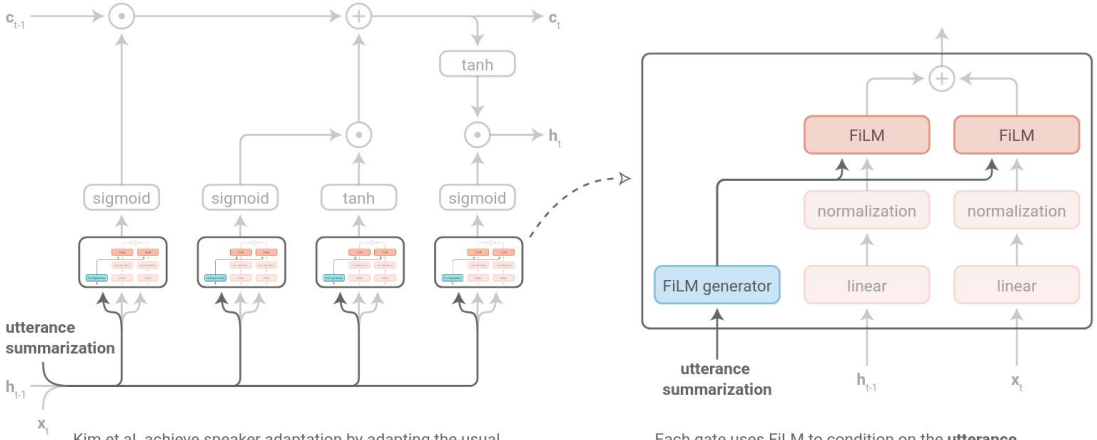


Fig. 1. A Squeeze-and-Excitation block.

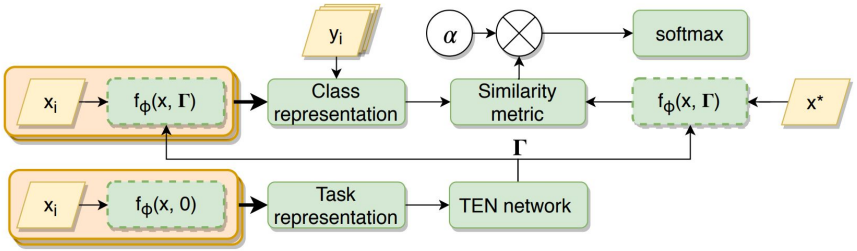
Dynamic Layer Normalization for Adaptive Neural Acoustic Modeling in Speech Recognition

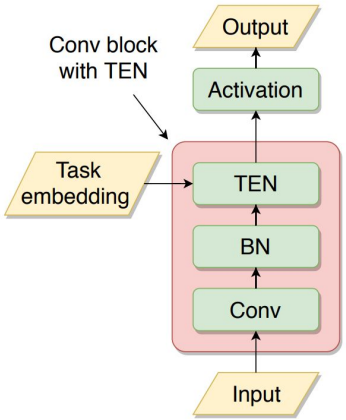


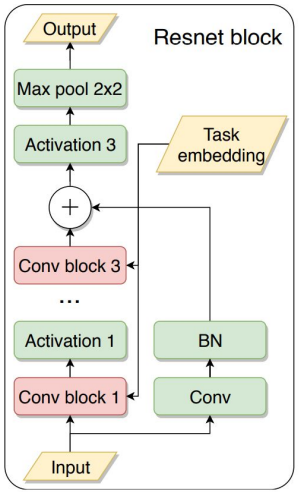
Kim et al. achieve speaker adaptation by adapting the usual LSTM architecture to condition its various gates on an **utterance summarization.**

Each gate uses FiLM to condition on the **utterance summarization**.

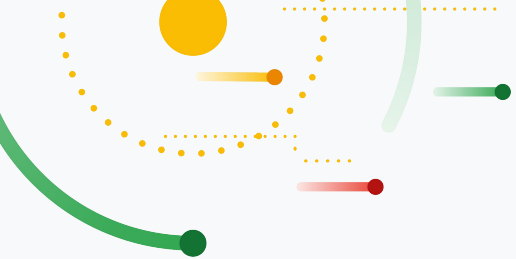
TADAM: Task dependent adaptive metric for improved few-shot learning







Oreshkin et al. (2018); figures from the article



Google AI

Properties

of feature-wise transformations

Task representation learning

FiLM parameters are an instruction on how to modulate **computation** in the task-solving network.

FiLM parameters are a **representation** for the task description.

Interpolations

Video produced by Dumoulin et al. (2017)

Interpolations



Animations produced by Theo Sanderson using a trained BigGAN model (Brock et al., 2019)

What are hidden layers?

A hidden layer is an abstract representation of the input.

A hidden layer is the intermediary state of a numerical program.

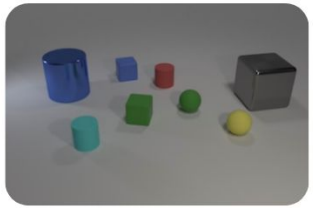
Task analogies

“king” - “man” + “woman” = “queen”*

* *Terms and conditions apply.*

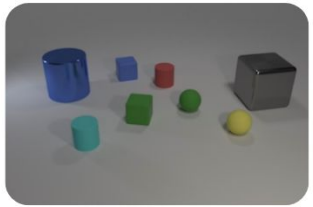
Task analogies

The model incorrectly answers a question which involves an unseen combination of concepts (in **bold**).



*Q: What is the **blue** big **cylinder** made of?*

Rather than using the FiLM parameters of the FiLM generator, we can use those produced by combining questions with familiar combinations of concepts (in **bold**). This corrects the model's answer.



Q_A : *What is the **blue** big **sphere** made of?*

Q_B : *What is the **green** big **cylinder** made of?*

Q_C : *What is the **green** big **sphere** made of?*

Figure from Dumoulin et al. (2018)

Learning and composing numerical primitives

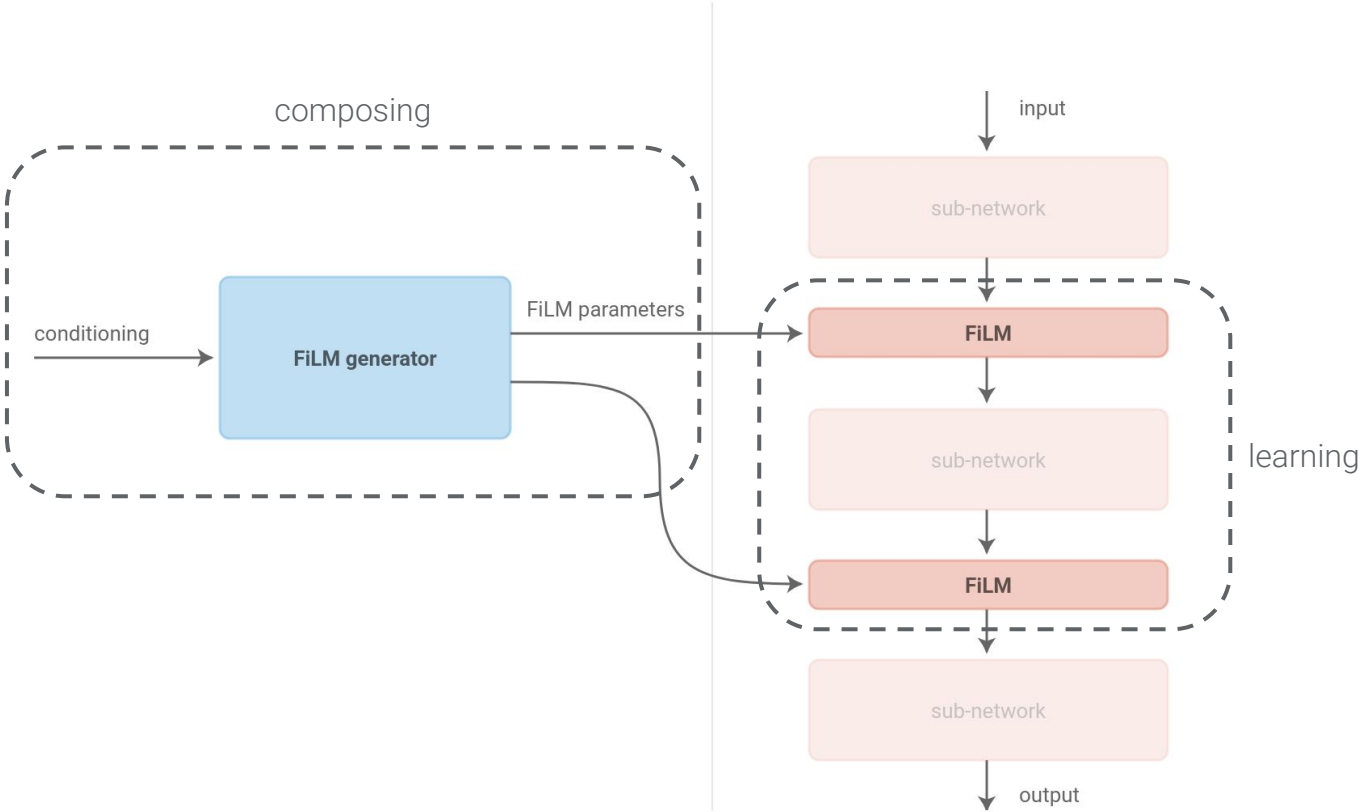


Figure from Dumoulin et al. (2018)

Interpretability

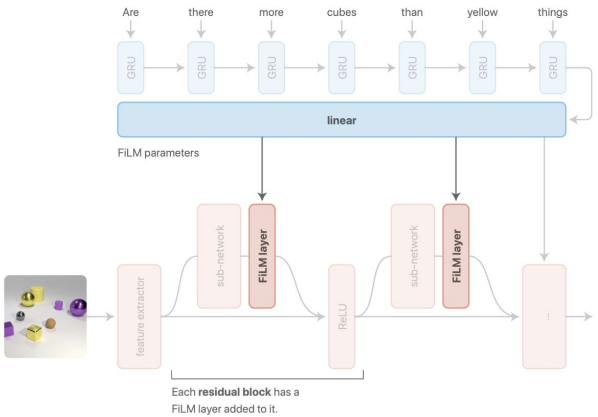
How can so few simple interactions compound into meaningful modulations of the task-solving network?

Interpretability

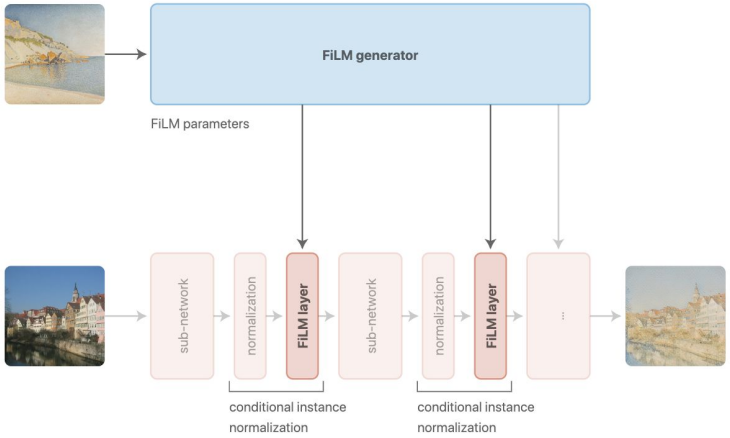
Is FiLM a selection mechanism for computational primitives?

Interpretability

Visual reasoning



Style transfer

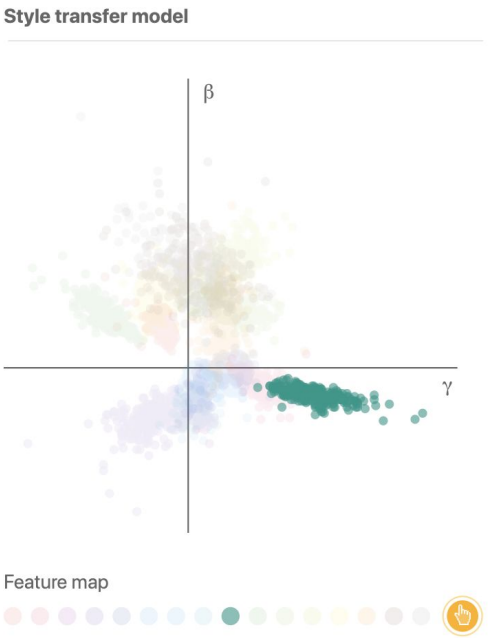


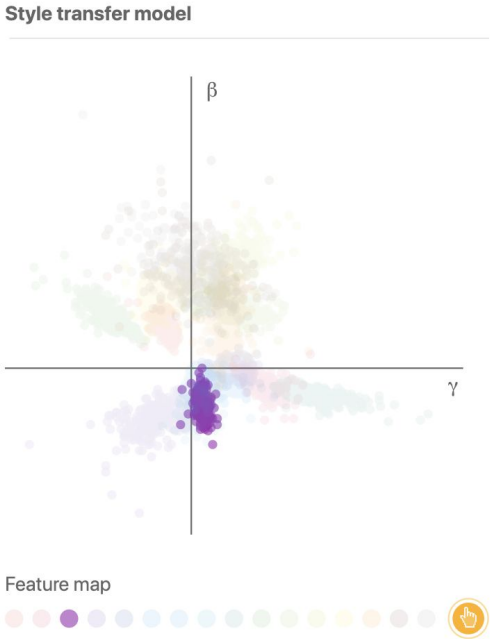
Figures from Dumoulin et al. (2018)

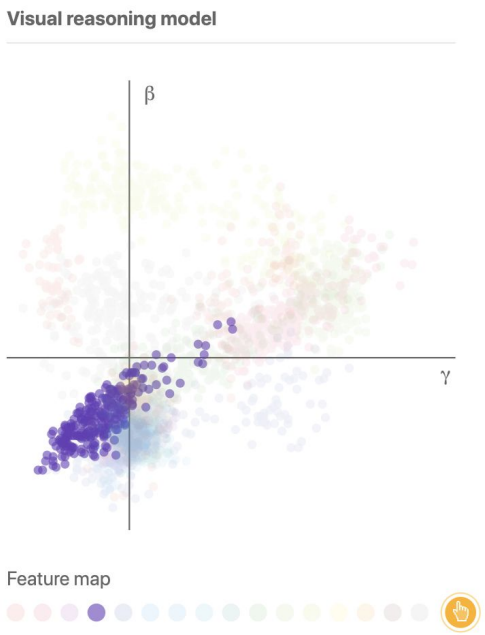
Interpretability

Figures from Dumoulin et al. (2018)

Interpretability





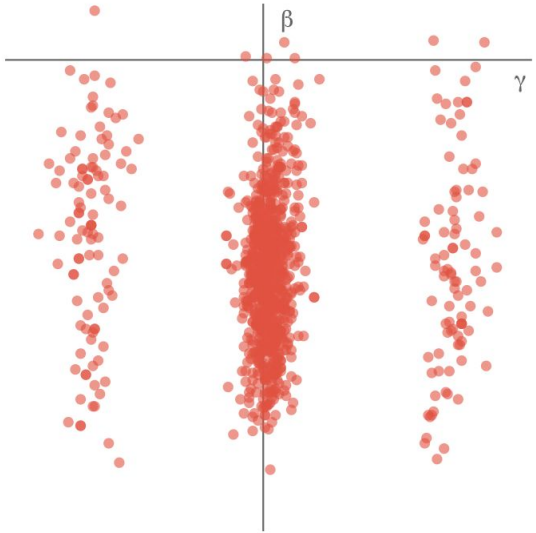


Figures from Dumoulin et al. (2018)

Interpretability

FiLM parameters of the **visual reasoning model** for 256 questions chosen randomly.

Feature map 26 of the first FiLM layer.



Feature map 76 of the first FiLM layer.

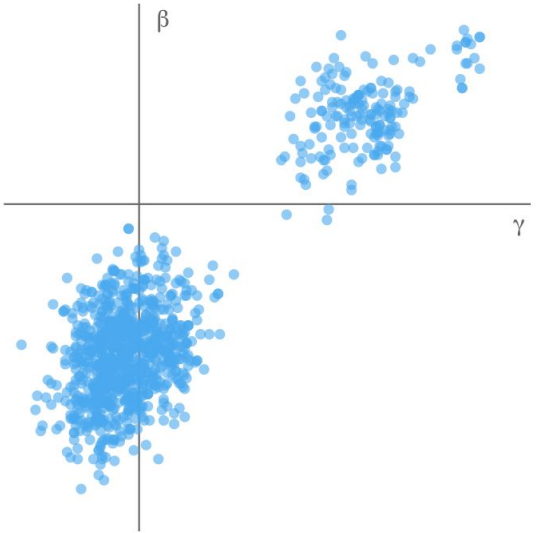


Figure from Dumoulin et al. (2018)

Interpretability

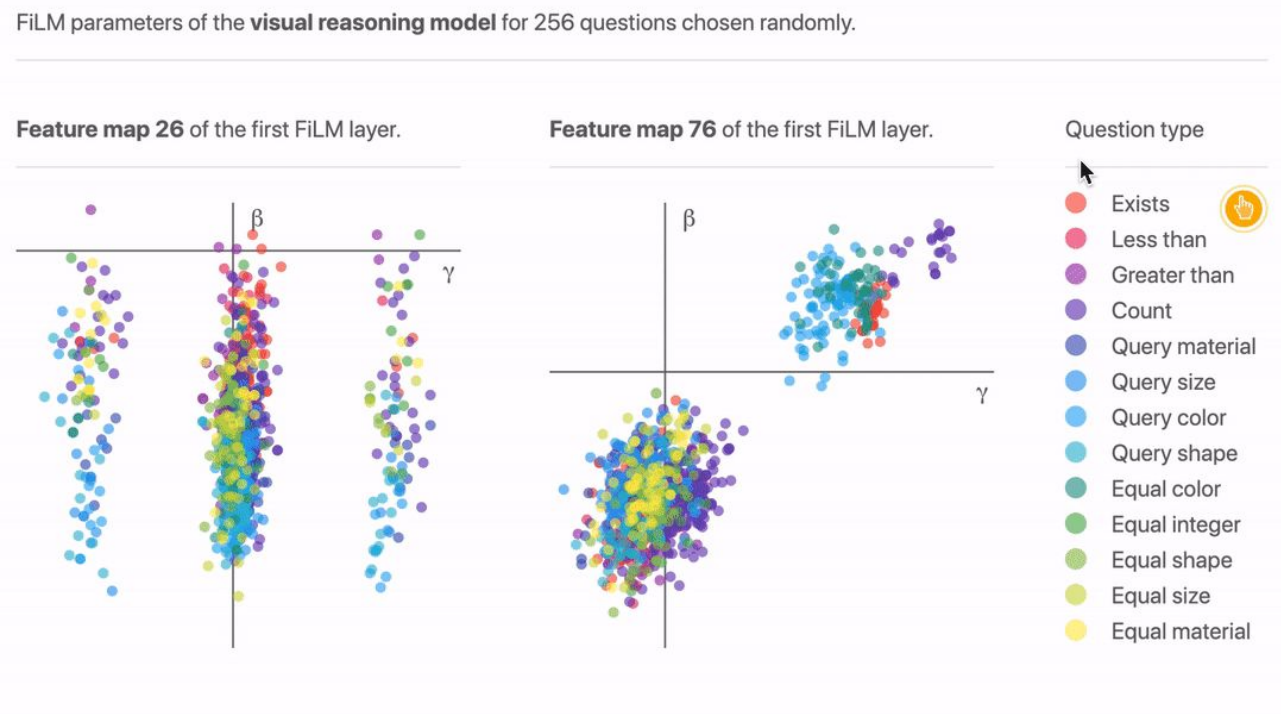
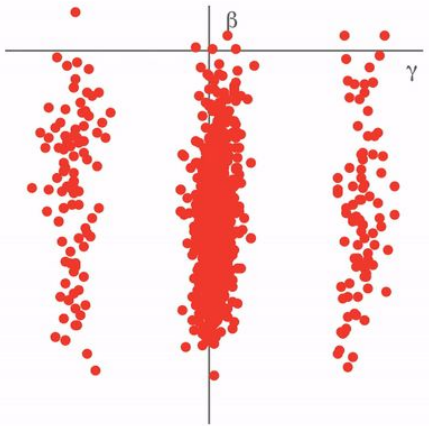


Figure from Dumoulin et al. (2018)

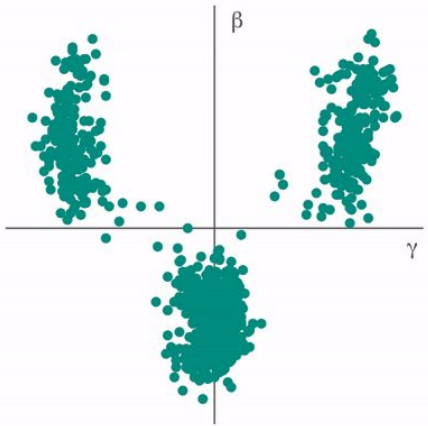
Interpretability

FiLM parameters of the **visual reasoning model** for 256 questions chosen randomly.

Feature map 26 suggests an object position separation mechanism.



Feature map 92 suggests an object material separation mechanism.

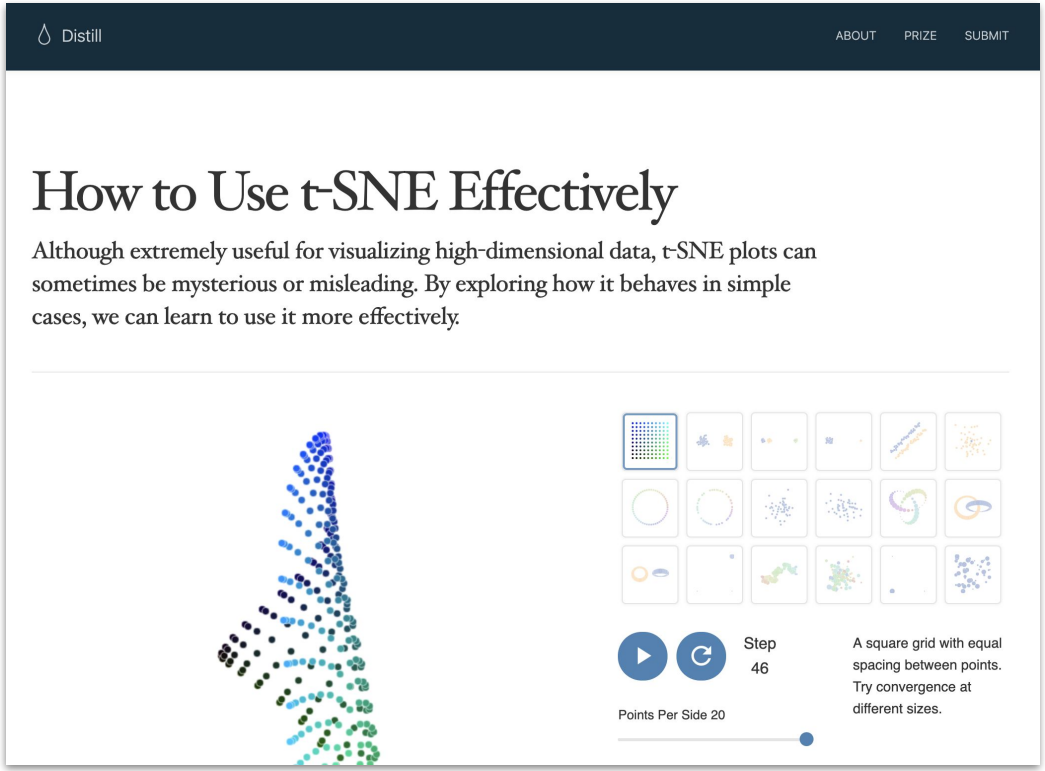


Word in question

front
behind
left
right
material
rubber
matte
metal
metallic
shiny

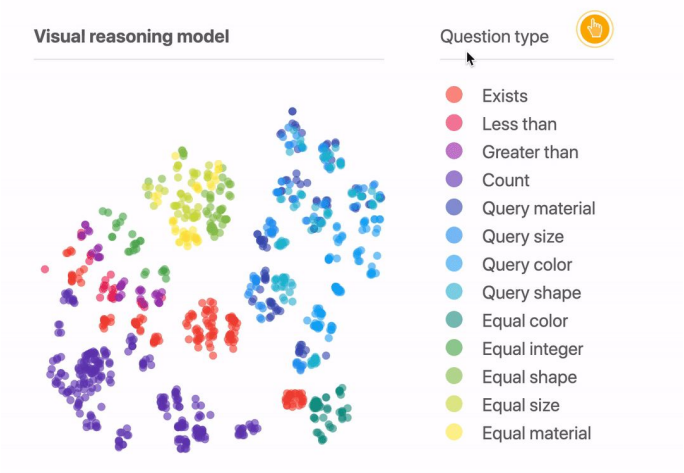
Figure from Dumoulin et al. (2018)

Interpretability



https://distill.pub/2016/misread-tsne/ (Wattenberg et al., 2016)

Interpretability



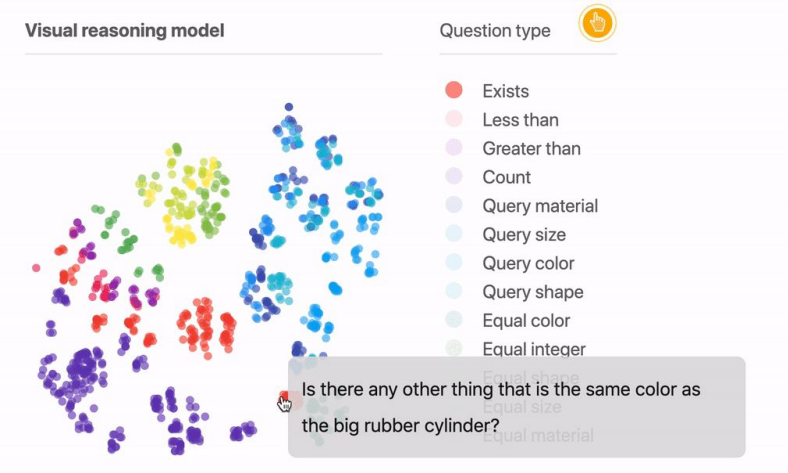


Figure from Dumoulin et al. (2018)

Interpretability

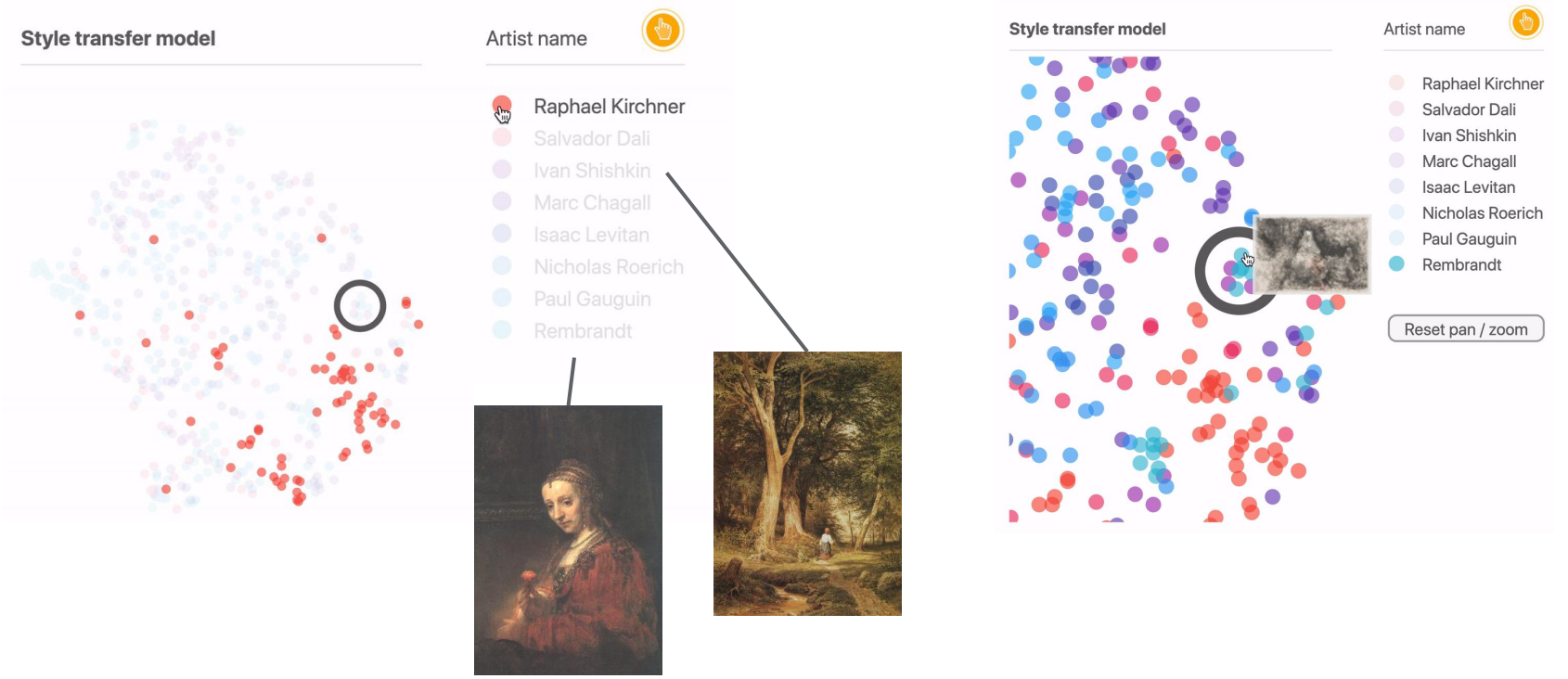
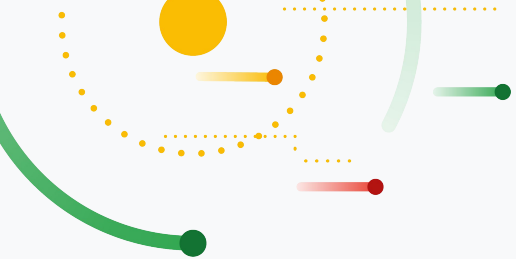
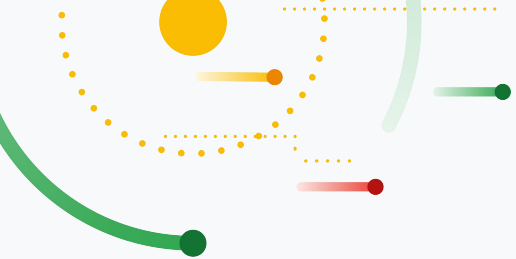


Figure from Dumoulin et al. (2018)



Google AI

Closing remarks



Google AI

Questions?

References

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of the International Conference on Learning Representations*.

Chaplot, D. S., Sathyendra, K. M., Pasumarthi, R. K., Rajagopal, D., & Salakhutdinov, R. (2018). Gated-attention architectures for task-oriented language grounding. In *AAAI Conference on Artificial Intelligence*.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In *Proceedings International Conference on Machine Learning*.

Dumoulin, V., Shlens, J., & Kudlur, M. (2017). A learned representation for artistic style. In *Proceedings of the International Conference on Learning Representations*.

Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H. D., Courville, A., & Bengio, Y. (2018). Feature-wise transformations. *Distill*.

Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., & Shlens, J. (2017). Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *Proceedings of the British Machine Vision Conference*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

References

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kim, T., Song, I., & Bengio, Y. (2017). Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition. In *Interspeech*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Oreshkin, B., López, P. R., & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

References

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*.

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-SNE effectively. *Distill*.

Yu, J., Yang, L., Xu, N., Yang, J., & Huang, T. (2019). Slimmable neural networks. In *Proceedings of the International Conference on Learning Representations*.