Decoding Vision Transformers: the Diffusion Steering Lens

Supplementary Material



Figure 8. Resid post visualizations using Diffusion Lens [43] (every 4 layer)

Input							
~	E	(4)	S		5	Jone	
L3 CONKE GROE K. SRIP	CONCE GROVE K. SIRVEY	CONSE GROE K. SRIP	CONCE GROLE K. CIRE!	ODIKE GROE K. SRIP	CONCE GOOVE K. SINZ P	QOIGE GROE K. GRIP	OOKE GROK K. SINP Vira (data)
L7 RAME! 20F KE'SINRI MARY & ODAR	RAKE! 2017 ML'SMR!	RAISE / 20F XIL'STARI	RAKE 2011 OF STRAN	RAKE 201 XE'SDAR	RAKE 2017 XL'SARP	RAKE! 20F XL'GDAR!	RAKE! 201 XE'SDR!
L11 QAKE 2017 K. GORF	QAKE 20# N. GORF	QAKE 20F N. GORF	CARE 2011 M. GORT	QAKE 201F N. GORF	QAKE 2014 N. GORH	QAKSE 201F NL GOREF Renta & Dener	QAKS 20F K. GOR
L15 OABE 2006 K. SABR ward soor	OAHE 2016 K. SABH	OALIE 20E K. SABE	OABE 2016 K. SABE	OAHE 200E K. SABH	OARE ¹ 20F K., GOBF waterd const	OAHE 20EK SORF	OANE 2016 K. SABI
L19 DARE 2014 K. SARI ORDRAEG	QARE 2011 K. SABY	CANNE 2017 K. SORP	ONNE 2019 K. SARI	QARE 2011 K. SART	QAME 2011 K. SARY Yobdbabs	CARE 2017 K. SARY	GARE 2017 K. SABH
L23 Refer drop mak-d-ge	MBETY/ XOF MAR GGRI	HNIKLY XXF Mar Gor	MONE SAK <u>M</u>	RREPLOYON MAK-O-GR	QPIGE† 201E K. SOIRT	MBER KOR MAK- GGR	RABE P. GROW NAR- 0-64
L27 RARE? 2017 28 SORN URGRALD	MARCH ZRI MOR GOKI Rokkeb	OAHLMZGE K. SORP week odde:	kinge gadne kep garin	RAHEK 201F K. SOIRT LADKREG	QARE' 2011 K. SCR ROKKIG	Rakie? 2011 X. GCIR! Rokkeg	RAKE ¹ 2011 K. SCIR1 ROKKED
L31 NUKL D SIGI PAR GOKI	Plocki & Soci Kelis Soler	Makik Zar Hoert Sko D3iang	OORGE RANDHEDDL* SHRIN Dograe	NOKLIK Zett Rubs Gok	GONGE AUZOIF 207 GINRP	- 200	HOKUG JOI HAR GGKI
L35 DOINE ORHOR SILM		cokre gron sal' gron	Relat Dal Roba Soci		BHRLE-2RI PKR GOK	RMRUE ZRI RAZR GOKI	QOHE GROR KE GIRIN Vering (2017
	ROLIGOLIK GEN	Rescur - Show Majik . W- 205	ternel	NEAD-IKR 220R		RO PE CJOIF K. SHRN Kannek bood	olvinge köröff X. Ginin
L43	a	R. J.					
L47	REPARATE IN MARK		OOEKE JORON GAZI SIMEN Maeadki astuji	MRK ILIZZERT PARTIE GOR over & bowe	KIHLAVO K.K. 2000 LIR LARD & DE Opp. Game	DARLA 7 👼 -1.20 VoADKLEF	NIIMIR GAGH Girndi r Rya g th gas

Figure 9. *MLP out visualizations using Diffusion Lens* [43] (every 4 layer)

L47H14: 1.3e-02	L47H6: 8.3e-03	L35H15: 6.3e-03	L36H4: 5.9e-03	L46H12: 5.4e-03	L41H7: 5.4e-03	L29H2: 5.3e-03	L47H0: 5.2e-03	L45H7: 5.2e-03	L47H15: 5.2e-03	Input	Decoded Output
9-5	MERCE CONT	MGROK SEAR			Keen for its Dames		68D3500	NOV 7. Internet datase	Rok Dajor.o		
L47H6: 8.5e-03 00K6E 4620H 20 GBBN	L46H1: 7.3e-03 (KIGEA 20# 26. SIKBP	L36H4: 7.3e-03 ©01KE! 201F 2L." SMRH	L47H15: 6.7e-03	L46H11: 6.4e-03 DOIKE G201E K. SKIPP	L44H12: 6.3e-03	L46H7: 6.1e-03	L46H3: 6.1e-03 CARE? 200 (%. SARF	L47H10: 6.0e-03 ROINE? 2017 21. SAIRI	L47H1: 5.9e-03	1005-11- 300-10	
Sa E	or Babbo			à	GN DOIDED Internet to the and their added calls in the task and a		Age of it down	Kos Dšiono	MOK Dasa conce		A
L47H1: 7.1e-03	L47H15: 6.5e-03	L37H3: 6.3e-03	L45H1: 6.3e-03	L35H4: 6.1e-03	L45H9: 6.1e-03	L47H11: 6.1e-03	L45H12: 6.1e-03	L46H0: 6.0e-03	L46H5: 6.0e-03		
Kon Caloito	Xon Dáisto		CITAL ZON A. CONS.	Xoe Calobo	Xen Daiono	Xoo S Jiono	BD	GADKKII	BDöbto		
L46H0: 7.5e-03	L33H7: 7.3e-03	L0H8: 7.2e-03	L45H7: 7.1e-03	L41H7: 7.0e-03	L47H4: 7.0e-03	L32H4: 6.9e-03	L0H13: 6.9e-03	L45H5: 6.9e-03	L25H4: 6.8e-03		
Zein Gzzef Guze Kok	QAKE! 201F KL SMR!	20KLD ZRI MAN GOR!	Ren 221 GsaicKi	Qoime grove Kid Ginan	MHRID ZRI MGR GOKI Kos Daiopo	QAKIE 2019 K. SORP	HORKLD XRI PAN GORI	MKIKOK GALIZA LIR2.IZIKGN 1977 2074 Galang 1971 Core Galan	QAME 201F K. SOIRF		
L47H15: 9.4e-03	L46H0: 5.6e-03	L44H9: 5.5e-03	L46H6: 5.5e-03	L46H12: 5.4e-03	L41H12: 5.4e-03	L37H7: 5.3e-03	L47H10: 5.3e-03	L37H5: 5.3e-03	L36H4: 5.3e-03		
MRIG 24H PRAPAGER	OORIE? 201F K. EIRBI Ken Dason d	GADKALY?	DANE JOH KL SIREF	ODIKE GROLE IO. SOEN	QARIE 2017-01, SMRH Kom Calintro	QAMEA 201 KB SCIEP BOKKEY	Nogkz & Daciak GRORNE	RANE* 2011 K. GMBI	QOIKE 201F2E SINRY		
L47H14: 1.2e-02	L47H6: 7.6e-03	L47H15: 7.3e-03	L46H0: 6.7e-03	L36H1: 6.5e-03	L46H11: 6.3e-03	L36H4: 6.3e-03	L35H15: 6.3e-03	L44H3: 6.3e-03	L32H11: 6.3e-03		
F.	ZOKCD ZIRE RAIN G OR	DUINER, 201 201 GRAP	UX H ARUE K. SOIP	UDINE / ZUH /2. GARAF	QUINE GOUF AL SUNRY		GARLY ZON DE SUEN Kom Daiot.o ^r	Kom Daiot:o	CARE ZUP K. SUIK	4	5
L47H14: 1.2e-02	L46H9: 8.8e-03	L46H12: 8.6e-03	L36H4: 8.1e-03	L46H11: 7.8e-03	L47H6: 7.8e-03	L45H12: 7.8e-03	L45H9: 7.8e-03	L37H3: 7.7e-03	L45H8: 7.6e-03		
	ନ = ୮.	WHAT GOOD 25 SUBL			KMF A-BIKY AZ -SEK CK GE nr 14 Qen 0 20220	идме 2017 К. БОВР овДарто	V68OMKEG	Barrink Good	GAMMELD ZRE KOR GOK	Jan St.	
L46H0: 7.8e-03	L46H7: 6.8e-03	L36H4: 6.7e-03	L45H12: 6.5e-03	L36H2: 6.4e-03	L36H9: 6.3e-03	L34H12: 6.3e-03	L38H6: 6.2e-03	L47H4: 6.2e-03	L37H15: 6.1e-03		
LIR TAIR. Moar Rkgok	Kek Gzlot Relir Golk		RDobco	Controlls" ZUN PA. GINDE	REFERENCE * ZUNY ZA. GUNBER	Barn for & Diment	en tolk * 2011 N. GUIRF	«ADRAEG	General 2006 No. GUNEY		

Figure 10. Top 10 heads in similarity with input when using Diffusion Lens.



Figure 11. Top 10 heads in similarity with input when using Diffusion Steering Lens.



Figure 12. Images with overlays.



Figure 13. Reconstruction results of images with overlays.



Figure 14. Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from clock-overlayed images.



Figure 15. Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from car-overlayed images.



Figure 16. Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from santa-overlayed images.



Figure 17. Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from fence-overlayed images.